DOCUMENT RESUME

ED 467 428 TM 034 314

TITLE	The Full Measure: Report of the NASBE Study Group on Statewide Assessment Systems.
INSTITUTION	National Association of State Boards of Education, Alexandria, VA.
PUB DATE	1997-10-00
NOTE	47p.
AVAILABLE FROM	NASBE Publications, 1012 Cameron St., Alexandria, VA 22314 ($12 plus $2 shipping and handling). Tel: 800-220-5183 (Toll Free); Web site: http://www.nasbe.org.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	*Academic Standards; *Accountability; *Educational Change; Elementary Secondary Education; Program Implementation; *State Programs; *Testing Programs

ABSTRACT

Nearly every state is engaged in what is generally being called "standards-based reform," the development of statewide academic goals and guidelines in core subjects, with new state tests that will measure students' progress toward these goals. The first chapter of this report describes how the Study Group on Statewide Assessment of the National Association of State Boards of Education began considering the testing aspect of standards-based reform, and then proposes two recommendations: (1) a state's assessment system should be designed to ensure that all students have the opportunity, assistance, and incentives to meet the state's academic standards; and (2) state assessments should be only one component of a larger system of evaluation and accountability that also includes system evaluation through multiple indicators, professional development, student remediation plans, and public engagement. The second chapter discusses elements of an effective state assessment system and outlines the characteristics that will be necessary. The third chapter discusses implementation issues for state assessment issues, including linkages with higher education and professional education and the costs of such assessment. Four appendixes contain a glossary; National Assessment of Educational Progress reading frameworks; standards for program, project, and material evaluation; and a list of consultants to the study group. (Contains 54 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 428

NASBE
NATIONAL ASSOCIATION OF STATE BOARDS OF EDUCATION

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

B. Welburn

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

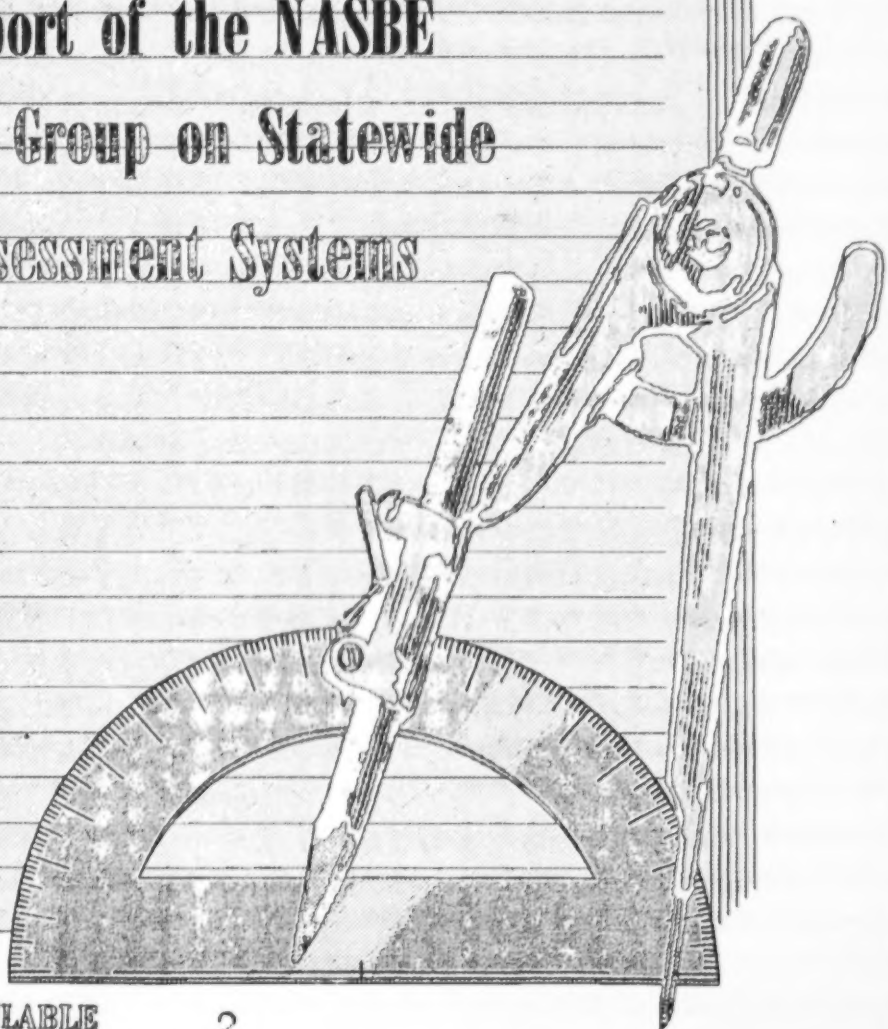
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM

The Full Measure

Report of the NASBE Study Group on Statewide Assessment Systems

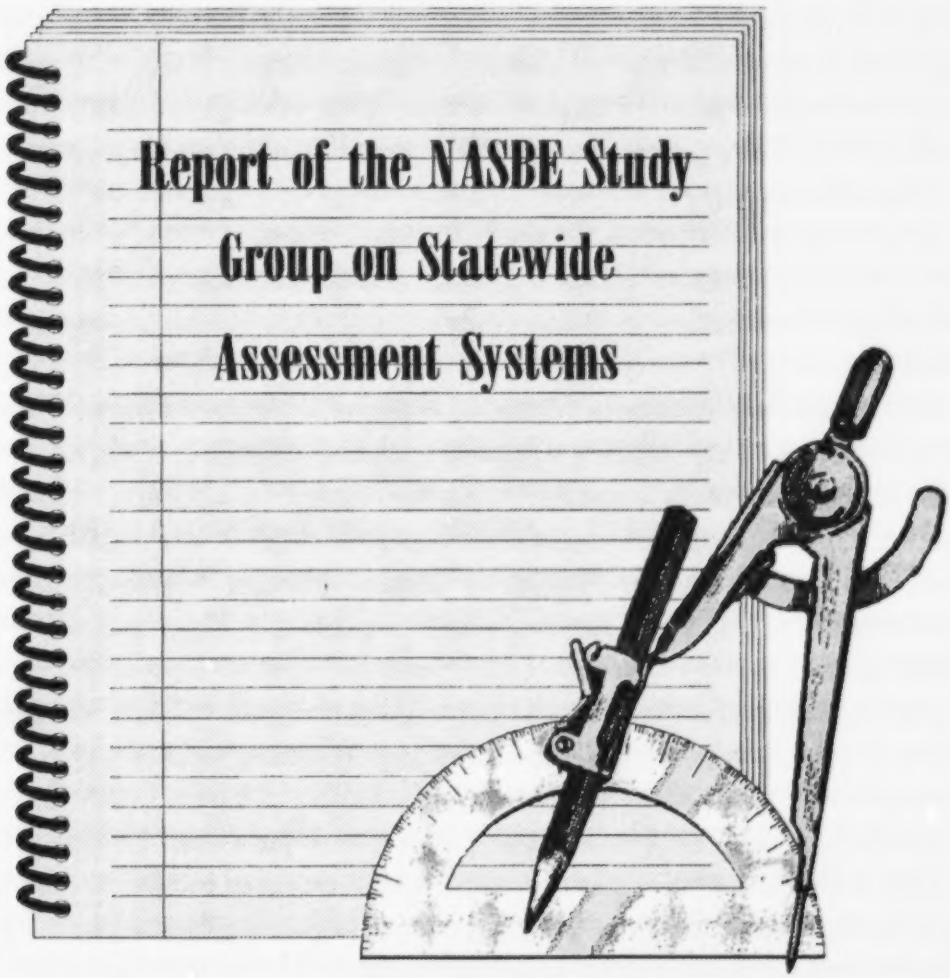


TM034314

BEST COPY AVAILABLE

NASBE
NATIONAL ASSOCIATION OF
STATE BOARDS OF EDUCATION

The Full Measure



October 1997

National Association of State Boards of Education

Study Group on Statewide Assessment Systems

Martha Wise (Ohio), Chair
Laken Cosby, Jr. (Kentucky)
Anne Dillman (New Jersey)
Edward Donley (Pennsylvania)
Kathryn Dronenburg (California)
Margaret Erling (Oklahoma)
Wendell Maddox (Minnesota)
Marlis Mann (New Mexico)
Kay McDonough (Utah)
James McLarty (Arkansas)
Talmadge Portis (Mississippi)
Jane Norwood (North Carolina)
Roberta Schaefer (Massachusetts)
Susan Stitham (Alaska)
Kathleen Straus (Michigan)
Sharon Williams (Missouri)
Kathy Wilmot (Nebraska)
John Wisthoff (Maryland)
Jeffery Zaring (Indiana)

NASBE Staff

David Kysilko
Carla Claycomb

The Study Group would like to acknowledge the valuable support and expertise it received from the members of its Advisory Panel, who are representatives of a number of the leading organizations associated with test development. Members of the Advisory Panel met with NASBE President Tom Davis and several NASBE staff prior to the first meeting of the Study Group to help define the essential issues and questions that would underlie the Group's work. Each member also had an opportunity to present to the Study Group information on a particular topic related to assessment and to review and provide comments on a draft of the final report. Finally, members of the Advisory Panel generously provided funds to help offset some of the travel-related expenses of Study Group members and speakers.

NASBE is very grateful for the assistance provided by the Advisory Panel and their efforts to work with state boards as partners in improving educational achievement. However, the recommendations and findings included in this report are solely those of NASBE and the Study Group on Statewide Assessment Systems and do not necessarily represent the views of the Advisory Panel or any other organization or individual associated with the work of the Study Group.

Advisory Panel

Dr. Howard Everson, Chief Research Scientist, The College Board, New York, NY
Dr. Wayne Gressett, Vice President, Harcourt Brace, San Antonio, TX
Dr. Stuart Kahl, Vice President, Advanced Systems for Measurement and Evaluation, Dover, NH
Dr. Michael Kean, Vice President of Public and Governmental Affairs, CTB/McGraw Hill, Monterey, CA
Ms. Patty McAllister, Director, District of Columbia Office, Educational Testing Service, Washington, DC

© Copyright 1997 by the National Association of State Boards of Education. All rights reserved.
Additional copies of this report are available for $12.00 per copy, prepaid, plus $2.00 shipping and handling from NASBE Publications, 1012 Cameron St., Alexandria, VA 22314. For more information call the NASBE Publications hotline at 800-220-5183 or visit NASBE's web site at www.nasbe.org.

Table of Contents

Chapter I. Standards and Assessments: Key to the State Role in Education 4

Chapter II. Elements of an Effective Assessment System 8

A. An Effective State Assessment System is Aligned with Rigorous State Standards
B. An Effective State Assessment System Addresses Specific Goals and Purposes
C. An Effective State Assessment System Balances Validity, Reliability, and Efficiency
D. An Effective State Assessment System Informs Remediation and Has Consequences
E. An Effective State Assessment System Has Mechanisms to Encourage Schools and Districts to Use Tests That Are Aligned with the State System
F. An Effective State Assessment System has a Clearly Articulated Relationship with National and International Measures of Student Performance

Chapter III. Implementation Issues for State Assessment Systems 18

A. Communications and Assessments: Dealing with Politics, Public Engagement, Reporting Results
B. Including All Students in State Assessments
C. Professional Education
D. Linkages with Higher Education
E. Evaluating Assessment Systems
F. The Financial Costs of Assessments
G. Conclusion

Appendix A. Glossary of Assessment Terms 39

Appendix B. NAEP Reading Frameworks 40

Appendix C. Standards for Evaluation of Educational Programs, Projects and Materials 41

Appendix D. Consultants to the Study Group 42

Bibliography 43

CHAPTER I: Standards and Assessments: Key to the State Role in Education

For many years governors, legislatures, and state boards of education alike have struggled to find the most effective state role in improving student achievement and providing oversight for the education system. Today, nearly every state is engaged in what is generally being called "standards-based reform": the development of statewide academic goals and guidelines in core subjects, along with new state tests that will measure students' progress toward achieving these goals. This has proved to be a complex, lengthy, and at times frustrating undertaking. In focussing on the testing aspect of this new paradigm, NASBE's Study Group on Statewide Assessment Systems began with the most fundamental questions: Does standards-based reform make sense? What is new and better about the new standards?

What's New about "New Standards"?

In the past, American teachers and school systems generally relied on "implicit" standards to guide instruction. What a student was expected to know was basically 1) the knowledge contained in textbooks adopted by states, districts or schools; 2) the knowledge and skills used by large-scale, norm-referenced achievement tests — the ACT, SAT or Iowa Test of Basic Skills, for example, or any of various minimum competency tests states required students to pass in order to graduate; 3) the knowledge embedded in a course curriculum developed by individual teachers in a school or by district-level curriculum coordinators; or 4) some combination of the previous three.

In the 1980s, many began to believe that these implicit standards were too low — that textbooks were being "dumbed down," that grade inflation was allowing students to receive As and Bs for average work, that scores on national exams were falling or stagnant at best. By 1989 a combination of events, including the publication of curriculum standards by the National Council of Teachers of Mathematics, the development of national education goals at the national education summit in Charlottesville, Virginia, and continuing bad news on the student achievement front (especially for minority and at-risk students) began propelling many educators and policymakers toward a system of state or district academic standards based on high expectations for all students.

At the same time, another movement calling for a whole new outlook on education oversight began to emerge. This vision, borrowed from corporate America, called for state systems to reduce their efforts at regulating the "inputs" of education (everything from specifying the number of hours a school should devote to instruction in various subjects to the number and kinds of books that should be in a library) and instead focus on the "outputs" of the system, particularly the performance of students. Among other things, this would mean de-emphasizing the accumulation of Carnegie units or course credits as a measure of success (since no one could be certain that passing a history class in one school or district was the same as passing a history class in another) and instead looking at what students should "know and be able to do." The idea for states would be to set clear, definable goals at the top of the system and then give teachers and local administrators more flexibility to determine how to help students achieve them. These goals would be the state's content and performance standards.

What is "new" about today's standards, then, is that 1) they came to exist in a systematic way

that is publicly available for any teacher, parent or student to see; 2) they are set at high levels so that students are expected to learn challenging material (not just the basics), analyze information and communicate results; 3) they are applied to all students, not just the college bound; and 4) they are becoming the foundation of states' education systems. In short, the Study Group affirmed that for policymakers, educators and the public, a host of issues — including the importance of focusing on outcomes, providing flexibility to local schools and districts, the erosion of the Carnegie unit, variations in grading practices, student mobility, and the need for clear goals and targets for everyone — have combined to make standards "one of the most powerful options for school reform." (Marzano and Kendall, 1996).

The Importance of Assessments in Relation to Standards

If standards provide the skeleton or framework on which states can build their academic system, assessments are the muscles that bring the system to life and make it work. Most of the 48 states and other jurisdictions that are developing or have developed standards are also adjusting their state assessments to align with the new standards. Or they may be developing totally new assessment systems. Whatever the case, the nature of these new assessment systems may well determine whether the standards truly become the central goals and framework that drive teaching, learning, equity and accountability in the state or whether they are merely vaguely-remembered pronouncements that gather dust in districts' central office buildings.

There are several key factors that tie assessments to the success of state standards:

1) Assessments help assure that standards are taken seriously. When the content of state assessments is closely aligned with that of the standards, this is a clear signal to teachers, administrators, students and the public that the state is serious about its academic standards. Basically, as all teachers and students know, if "it's on the test" it must be important. Aligning assessments with standards is critical for another reason, as well: if material is covered on the assessment but is not in the standards, it is inherently unfair to students and teachers alike. In such cases, the material may not have been covered in class because, again, no one knew it was important.

Still, one more step is critical: not only must content be both in the standards and on the assessment, the assessment must mean something: in today's parlance, there must be "stakes" attached to the assessments at some point. For students that point is most likely to be in terms of a diploma: that is, a student's level of achievement on state assessments could be indicated on the diploma or meeting the state's academic standards at, for example, the 10th grade level, could be one requirement for graduation. For districts and schools, it means student scores on state assessments should be one component of the state's accountability plan.

2) Standards and assessments guide teaching and learning. Clearly articulated content standards and the kinds of examples used in performance standards or curriculum frameworks to illustrate the knowledge and skills embodied in the standards are intended to guide teachers toward the state's vision of essential teaching and learning. While some teachers may adhere closely to the types of examples given, others may work more from the "spirit" of the examples in creating their lessons. The point is that when standards are taken seriously through their embodiment in state assessments, all teachers will share a vision — shaped by a consensus of the best thinking in the state — of the knowledge and skills students should acquire.

3) Assessments help individual students meet the standards. The ultimate goal of state standards and assessments is not sorting students into the "haves" and "have nots" of knowledge. The goal is to enable all students to meet the standards. In this sense, results on state assessments should be part of a support system that can signal when a student is in trouble academically and needs extra help or special instruction. At the same

time, teachers should be able to use the standards in developing their own classroom assessments to help identify students who may need more instruction or to inform a teacher when it is time to adjust his or her curriculum or instructional strategies. If states are serious about having all students meet the standards, they will provide assistance to districts and schools for this remediation.

4) Assessments help policymakers ensure that all students have access to a sound education. Policymakers want to assure the public that all students, no matter where they live in the state, have a reasonable opportunity to get an education that enables them to meet state standards. As indicated above, having assessments as one part of the state's accountability system helps ensure that the standards are taken seriously. Just as important, though, is using the accountability system to verify that all schools and districts are providing a "reasonable opportunity." This is a basic state responsibility — indeed, it is often specifically addressed in state constitutions. Some have expressed concern that standards and high stakes assessments will only mean that more of those students who were having trouble in the past will not graduate in the future. Instead, the Study Group believes that standards and assessments, when part of an overall evaluation and accountability system, can help provide the state with the philosophical framework, the information, and the mechanisms it needs to bring a solid education within the reach of every student.

These four basic purposes of state assessment systems — and how they are linked to state standards — form the foundation of the Study Group's primary recommendations to state boards and other education policymakers.

Recommendation 1: A state's assessment system should be designed to ensure that all students have the opportunity, assistance and incentives to meet the state's academic standards.

Recommendation 2: State assessments should be only one component of a larger system of evaluation and accountability that also includes system evaluation through multiple indicators, professional development, student remediation plans, and public engagement.

The implications of these short recommendations are far reaching and will be spelled out in more detail in the ensuing sections of this report. But briefly, what it will mean for state boards of education and other policymakers as they implement these recommendations includes the following points:

- First, it means that assessments should be closely linked to the state's standards — and that the standards must contain a sufficient degree of specificity to provide guidance to those developing curricula and to those developing the assessments.

- Because today's standards are generally at a high level, asking both for specific knowledge and for thinking and problem-solving skills, it means assessments must be challenging, of high quality, and fully able to assess the knowledge and skills asked for in the standards. It means some portion of state assessments should be performance items that ask students to work with more complex issues and problems, demonstrate their thinking, and communicate their logic. The science of large-scale assessment has improved since the days when "teaching to the test" was almost universally a negative concept. The Study Group believes that teaching to a test can be acceptable — even desirable — as long as the tests have broad coverage of content and skills and specific questions are changed routinely.

- It means teachers must be able to use the assessments to adjust their course plans to cover areas where more work is needed or to identify individual students who need additional help.

- It means teachers themselves will need training in terms of understanding the new standards and assessments, what changes in curriculum and instructional practices will be

needed, and how teachers can incorporate state standards into classroom assessments.

- Because standards are designed for all students, it means that all students must be included in the assessments.

- It means the assessment system itself will be multi-dimensional: no one test or type of test can adequately accomplish all the purposes implied by these recommendations, and other components in addition to tests must also be part of the state's evaluation system.

- Because the assessments would have stakes attached for students, districts and/or schools, it means policymakers must be cautious yet firm in the implementation and "roll out" process. A time line that is too short, for example, could lead to lawsuits down the road.

- It means that assessments should be linked in some way to national and international measures of student performance, so parents can be assured that the state's standards are "world class;"[1]

- Finally, it means that the state's assessment system will not be especially cheap (though neither must it be a budget buster). Yet because standards, assessments and accountability represent the very foundation of the state role in education, it should not be otherwise. Policymakers cannot expect educators, students or the public to take the standards seriously if they are overly stingy in implementing assessments and accountability.

This, then, is the philosophical foundation for a state's assessment system. It is based on the fundamental state role in education: to **set academic and performance standards** for students, educators, and the "system"; to **monitor performance** in meeting these standards through fair, high quality assessments and collection of other data; and to **provide the necessary assistance and consequences** to ensure that all students have an opportunity to achieve the standards.

The Study Group fully understands that fulfilling this role will not be easy. It will require, at a minimum, ongoing adjustments to the system based on state experiences in this new way of operating and on continuing research and improvement in the science of assessment. At times it may require more serious undertakings, possibly including revisions of standards documents. And it may also require significant efforts on the part of politicians and policymakers to galvanize the support needed across the state in legislative chambers, in classrooms, in boardrooms, and in homes. Yet what is at stake is clear: state standards — indeed, the state role in education — may not be taken seriously and may not have the effects on instruction and student achievement desired by policymakers and the public unless these fundamental recommendations are considered. On the other hand, successful development and implementation of a high-quality state assessment system will demonstrate a commitment to improving education that the public demands and is willing to support.

---

1 Although making standards and assessments "world class" should be the goal of any state, the fact is that "world class standards" are very elusive and remain the topic of considerable debate. While learning from the curricular examples of school systems in other high-performing countries is commendable, the study group recognizes that the cultural, historical, and educational complexities involved in such comparisons are very difficult to do in a meaningful way.

BEST COPY AVAILABLE

CHAPTER II. Elements of an Effective Assessment System

States today have adopted a range of assessment systems with different consequences and purposes. The study group believes that diversity among state assessment systems is not inherently bad; on the contrary, they reflect the unique content and character of each state's education system. However, **the Study Group concludes that there are general characteristics highly effective state assessment systems share**. In general, an effective state assessment system is one that: is aligned with state standards; is designed to address specific goals and purposes; balances validity, reliability and efficiency; informs remediation and has consequences attached to some results; provides a framework for school and district initiatives; and has a clearly articulated relationship with national and international measures of student performance.

A. An Effective State Assessment System is Aligned with Rigorous State Standards.

Standards need at least two characteristics to be effective. First, they must be rigorous, explicit, and detailed enough to guide curriculum and assessment development. Second, standards must be put in force via a closely aligned assessment system. Many states have developed standards that are not rigorous or specific enough to meaningfully direct assessment development. In these states, it is likely to be the content and structure of their assessments, rather than the content of their standards, that drive school reform. Because assessments, more than standards, shape the behavior of students and teachers, standards can be undermined by unaligned assessments. An apt adage to remember here is:

- You get what you assess;
- You do not get what you do not assess; so
- Build assessments toward that which you want educators to teach (Resnick & Resnick, 1992).

B. An Effective State Assessment System Addresses Specific Goals and Purposes.

Whereas standards define what to test, assessment goals define who to test, when to test, and what type of test to use. The Study Group asserts strongly that the primary purpose for any state assessment system should be to improve teaching and learning. However, tests to improve teaching and learning can take many different forms and address a variety of complementary goals and purposes. For example, tests to inform student remediation need to be given to every student in the testing grade, constructed to give rich information to teachers (obviously, this implies that teachers must receive much more than student scores), given sufficiently early in the year to inform instruction, and designed for efficient scoring so that results can be returned to the classroom as soon as possible. Tests to evaluate schools or districts can be given to a sample of students and need not be given every year, which saves money and instructional time.

Aligning state assessments with state goals and purposes make good financial sense. **Assessments that are implemented without clear goals may be measuring variables that are unnecessary, or may be measuring necessary variables in inefficient ways.** For example, measuring every student's writing ability to inform district curricular improvement in writing is "overkill"; only a sample of students need to be tested for that purpose. Alternately, if a state goal is to ensure that all graduating seniors can write a well-formed paragraph, it is necessary to administer a writing exam to all seniors.

C. An Effective State Assessment System Balances Validity, Reliability and Efficiency.

For the results of any assessment to mean anything, the test must be both valid (meaning that it measures what it purports to measure) and reliable (meaning, among other things, that students would get similar results if tested again). Statistically, it is impossible for a test to be completely valid and reliable, but even making them "sufficiently" valid and reliable to satisfy stakeholders and be fair and equitable is a challenge.

Fundamentally, those who write assessments have two types of test questions to choose from: multiple-choice and performance. In general, multiple-choice tests are very reliable provided they include a sufficiently large number of items. Even though they require large numbers of items, multiple-choice tests are also efficient in terms of testing time and cost. But they do not resemble good instruction. Nor can they adequately measure how effectively teachers teach or students apply knowledge. Furthermore, multiple-choice tests essentially measure the ability to *choose* the correct answer, which is not a skill often required by adults in the "real world." For these reasons, multiple-choice tests have been criticized for having low validity.

In general, performance assessments are likely to measure skills that policymakers want students to develop, such as the ability to write a persuasive paragraph or to draw inferences from a science experiment. However, there are many types of performance assessment, each of which has unique strengths and weaknesses:

- Constructed response questions require students to write short essay responses or solve problems and show their work;
- Extended performance tasks require students to engage in individual or small group hands-on activities, and
- Portfolios or collections of student work collected over a long period of time, often include work students had time to review with others and revise.

Experience in many states, including Vermont, Kentucky and Maryland, indicates that the various types of performance assessment may differ in their validity and reliability.[2] Constructed response items administered in a traditional testing situation are more likely than other types of performance assessment to be able to be scored accurately enough to produce reliable test results. Of course, they are less efficient than multiple-choice questions because of the greater testing time per item, because more time and money are required to train individuals to score responses accurately and consistently, and because more time is needed to score student responses individually and by hand.

Like constructed response assessments, administering performance tasks tends to be very inefficient in terms of testing time. However, performance tasks model "curriculum imbedded assessment," meaning they more closely ask students to reproduce the kind of work done in class than other types of assessment. Although still controversial among more traditional assessment advocates, many believe that if administered properly the results of these assessments can accurately reflect the capabilities of individual students.

Portfolios can be effective methods to help teachers target instruction for student remediation or acceleration. However, when used as components of state assessments they have two primary problems. First, when students are allowed to get

---

[2] Another important but less talked-about issue is "system reliability." In this instance, even if the test itself is reliable, how the assessment system administers the test or how it uses the test results may cause problems. For example, states and districts that try to chart school improvement year by year or even over two-year periods should be concerned about how meaningful their information is given natural variations in student ability from, say, one 4th grade class to the next. As assessment researchers Baker and Linn (1996) write, "The fluctuations due to differences in the students themselves could swamp differences in instructional effects...."

BEST COPY AVAILABLE

help and revise their work, scores do not necessarily reflect students' capabilities. Second, when local scoring by teachers is the only way to make statewide portfolio assessment feasible, then scoring consistency across the state can be problematic. Some preliminary results from Kentucky demonstrate that over time teachers across a state can learn to score fairly consistently; however, auditing and training individuals to evaluate portfolios is very expensive and time consuming.

Because no one test can "do it all," **the Study Group recommends that states consider ensuring fairness, curricular coverage and depth by implementing an assessment system with more than one test that balances performance assessment items with more traditional multiple-choice.** The National Assessment of Educational Progress (NAEP), the nation's primary measure of what students know and can do on a state and national level, is a good example of a test that combines multiple-choice and performance items and has proven both valid and reliable for use at the state and national, rather than individual, level. Unfortunately, many states rely exclusively on multiple-choice assessments; excluding writing assessments, 26 states rely entirely or nearly entirely on multiple-choice tests to measure student knowledge and skills in all subjects (Neill, 1997).

D. An Effective State Assessment System Informs Instruction and Has Consequences.

Two primary purposes for developing a state assessment system are (1) to inform student instruction by pointing out student learning gaps and students needing acceleration, and (2) to assign rewards and sanctions to students, teachers, schools or districts based upon assessment outcomes.

POINTING OUT LEARNING GAPS. The Study Group asserts that once states establish a system of standards, they have both a moral and legal obligation to provide all students with the instruction they need to attain them. If a student's performance on a state assessment points out knowledge deficiencies, the state has an obligation to assist all students in meeting content and performance goals. Assessments are instrumental to diagnose deficiencies and help teachers target instruction for individual students. Especially in elementary grades, a state test can

What makes a successful test with high stakes consequences?

When important rewards and sanctions are assigned in part according to assessment results, the test components and decisions that are based on them must meet rigorous standards for validity and reliability; test only those topics that students have had a fair opportunity to learn; and be free of racial, ethnic or cultural bias. High stakes assessments that do not meet these criteria are open to public criticism and legal challenges. Joan Baratz-Snowden of the American Federation of Teachers asserts that any successful high stakes assessment must have ***"APPLE"*** characteristics:

- **A**dministrative feasibility — the burden of administering the assessment, in terms such as lost instruction time and ease of administration, must not be prohibitive;
- ***P***rofessional acceptability — teachers and principals must support the assessment;
- ***P***ublic credibility — the public must believe the assessment is fair and measures important skills and knowledge;
- ***L***egal defensibility — the test must be valid and reliable and not have disproportionately adverse effects on students from particular racial or ethnic groups; and
- ***E***conomic affordability — the price of the assessment has to be figured sensibly and proportional to the state education budget.

be a powerful tool to focus attention and resources on helping students, schools, and districts meet standards. Unfortunately, few states are including remediation in their assessment agendas. According to the AFT(1996), 13 states require or have plans to require remediation for students not meeting standards, but only ten backed the requirement with funds. For example:

- In Ohio, the state requires that districts provide remediation for all fifth-graders who fail a fourth grade assessment in one or more core subjects. The state provides districts with funds for this, and included in each state curriculum framework is a section of suggested intervention programs.

- New York requires districts to provide remediation to any student who fails any reading, writing, or mathematics assessment in the elementary grades. Districts must also notify each student's parent or guardian of the student's test results and of the plan for the student's remediation. Funds are made available by the state for these remediation activities.

- Indiana has a comprehensive program of mandatory remediation. Students who perform in the two lowest categories (both below passing) on state assessments in reading and math must receive remediation. Students who score only slightly above or below passing (the next lowest categories) are also eligible for assistance. Indiana's funding formula for remediation directs money to schools with the most students scoring in the lowest two categories, and schools can apply for funds and services to serve students scoring in the next lowest categories.

INFORMING ACCELERATION. Helping all students reach their capacity requires that teachers receive high quality information about students who achieve highly on state assessments as well. **The Study Group asserts that once states establish a system to encourage all students to achieve high standards, they have an obligation to provide high achieving students with the instruction they need to develop their full capacity.** If a student's performance on a state assessment points out subject areas in which students have far exceeded state expectations, state policymakers, districts, schools and teachers need to have mechanisms in place, based upon assessment results and state standards, to encourage that student to continue achieving beyond state expectations.

HAVING CONSEQUENCES. Although providing remediation and acceleration to particular students is necessary for an effective accountability system, it is not sufficient. To have the greatest impact on teaching and learning, an assessment system also needs to have consequences. Making assessments count as one component of an accountability plan is important because studies show that high school students perform better on assessments that have consequences for them. Teachers are also more likely to take test results seriously and more inclined to change their instructional practices to reflect the content and format of a test if the test has high-stakes consequences (Truby, 1997)

Currently, sixteen states plan to tie the award of high school diplomas to achievement of their standards in core subject areas (AFT, 1997). In these, the assessments are commonly used in isolation, rather than as only one component of a multifaceted evaluation system. Denying diplomas to students based on test scores, when students are otherwise qualified to graduate means that students who do well in school but perform poorly on the state assessment may be unfairly penalized by a one-shot evaluation of their accumulated school work. Consequently, the Study Group asserts that either: 1) high stakes tests should be only one of several components that are considered when making decisions about student promotions and graduation; or 2) state assessments should be used in high stakes ways other than graduation or promotion, such as for conferring diplomas with endorsements.

More common than using state assessments to assign consequences to individual students is using assessment results for school or district accountability. According to the National Educa-

BEST COPY AVAILABLE

Legal Defensibility

Susan Phillips, a professor at Michigan State University who specializes in legal/policy issues in assessment, has identified a number of areas related to high stakes testing that could provide grounds for a legal challenge. These include:

- **Adverse Impact**. In testing, adverse impact would occur, for example, when "substantially more historically disadvantaged students fail a graduation test than nondisadvantaged students." However, courts have established that adverse impact "is a necessary but not sufficient condition for establishing intent to discriminate."
- **Due Process**. According to Phillips, "substantive due process can be violated when the knowledge and skills being assessed are judged to be arbitrary, capricious, or unfair." In addition, students must have adequate notice of the implementation of a high stakes assessment. Periods of less than two years have generally been declared by courts to be inadequate.
- **First Amendment Issues**. First amendment rights may be violated "to the extent that [standards] prescribe or proscribe certain viewpoints" or if there is the "suggestion that specific attitudes or values are being assessed."
- **Students with Disabilities**. Failure to provide reasonable accommodations may violate the rights of students with disabilities under the Americans with Disabilities Act (ADA). Phillips notes that the "most important requirement when balancing inclusion of the disabled against the validity of the assessment is to develop a comprehensive written policy outlining the procedures for requesting accommodations and detailing how decisions will be made...."
- **Opportunity to Learn**. The state or district must be able to "demonstrate that students have had an opportunity to learn the knowledge and skills included" on a high stakes assessment. Opportunity to learn (OTL) has been a difficult issue for both politicians and the courts. Many observers believe that further court actions will be needed to define the criteria by which states will be able to measure OTL. (Phillips, 1995)

tion Goals Panel (1996), 40 states use assessment results to inform some mechanism of school accountability. However, the same caution applies to schools and to districts as to individual students: no school should be penalized on the basis of test results alone. Rather, test results should be one component of an evaluation system that may also include variables such as dropout rates, teacher attrition rates, expenditure information and attendance rates.

The Study Group strongly recommends that states reconsider the acceleration, remediation and accountability components of their assessment systems to guarantee that students have the remediation or acceleration necessary to learn to the highest standards and that consequences exist that encourage everyone to achieve.

E. An Effective State Assessment System Has Mechanisms to Encourage Schools and Districts to Align Their Instruction and Evaluation with the State System.

Standards and assessments should serve as a catalyst to direct local reforms in improving teaching and learning. The assumption on the part of most states seems to be that instituting standards and assessments will prompt school districts to align their practices with them. Research suggests that such an assumption may be warranted for forward-thinking districts that have the capacity to tackle change head-on. Schools that lack the will or capacity to change, however, are unlikely to benefit from state mandates alone.

Explicit programs linking state standards and assessments with district and classroom teaching and assessment practices are critical to promote student achievement in all schools and districts. Research in Vermont, Arizona, North Carolina, Kentucky and Maryland show that one effective way to

impact classroom teaching and assessment practices is to include performance items on the state assessment. In general, teachers are more likely to incorporate the underlying goals and skills of a performance assessment in their instruction, but they need help to do so effectively. Several states have also developed other innovative programs to encourage teachers and superintendents to align local curricula and assessments with state standards:

- New Mexico requires that districts design their own reading assessments for students in first and second grade, the results of which must be reported to the state.
- Utah administers a norm-referenced standardized test in grades 5, 8 and 11. However, the state also offers districts a series of criterion-referenced and performance assessments to measure student achievement that is based on state curriculum frameworks.
- In Oregon, students in grades 3, 5, 8, and 10 must take statewide standardized tests and local assessments that are based on content and performance standards. While the statewide assessment includes multiple-choice, essay and problem-solving questions, the local test includes classroom assignments and other tasks that are difficult to assess in a standardized manner. These tasks vary among teachers and schools, but all students must complete a minimum number of specified tasks and achieve a passing score.

High Stakes Warning

While the Study Group believes that having consequences is an essential element of a state assessment system, it also emphasizes that attaching high-stakes consequences to assessment results is not a panacea for improving student achievement. At the same time, policymakers must consider several dangers to implementing a system of high-stakes assessments. Analyses of the effects of high stakes testing demonstrate that curriculum often narrows to little more than test content and teachers spend an inordinate amount of class time "prepping" students for the tests. Teachers, students and principals are also more inclined to cheat when sanctions are imposed according to test results. High stakes assessments can also spell trouble in terms of equity, since certain subgroups of the population, such as students in high poverty areas and many minority students, tend to score lower on standardized tests than state or national averages. These students may be especially vulnerable to a narrowed curriculum as they are drilled to improve test scores but receive little else in terms of a full, rich course of study. Finally, consequences of high stakes assessments could include increased rates of referral to special education (in particular when special education students are not included in testing or when their results do not have stakes attached to them) and increased rates of retention in grade (when only certain grades are tested), since retaining students excludes them from the test for another year after which they may perform better.

There are ways to guard against the more excessive negative outcomes of a high-stakes assessment system. Most fundamentally, the Study Group recommends that assessment results never be used in isolation to administer rewards and sanctions. Rather, test results should be only one component of an accountability system that includes a variety of other measures of student knowledge and skills. States may also consider changing assessment questions and content often to ensure that "teaching to the test" requires teaching a whole curriculum. Finally, states need to monitor the outcomes of their assessment system, to ensure that particular districts, schools or students, especially those from inner cities or from ethnic minority groups, are not unfairly penalized by the assessment program (see "Evaluation" section on page 35).

BEST COPY AVAILABLE

- North Carolina's "Next Century Assessment" will build local capacity by making 75 performance tasks and scoring guides available through the World Wide Web for teachers to use as instructional and assessment tools. Teachers are required to assess each student with tasks selected from this database, one of which is a common performance task required of all students in the state and used to calibrate teacher scoring of student work against state standards.

The Study Group encourages states to build a systemic system of school improvement that includes curricular, assessment, and other links between state policy and classroom practice.

F. An Effective State Assessment System has a Clearly Articulated Relationship with National and International Measures of Student Performance

As states invest time and resources in developing a comprehensive state assessment system, they must ensure they have some way to calibrate results on their assessment with the performance of other students, both nationally and internationally. Currently, most states that have articulated standards have designed assessment programs unique to their state. Because of their unique nature, states cannot compare their performance or that of individual students, schools or districts with those of other states or nations and have no quantifiable way of demonstrating that their students perform to national or global standards. **The Study Group contends that being serious about standards and assessments means ensuring that standards define, and assessments measure, world class performance.** But the majority of states neither compare nor calibrate the content of their standards and assessments with those of other states or nations. This is problematic not only because these states lack the ability to place their assessment results in national and international contexts, but because results from unaligned assessments can be confusing and contradictory to policymakers and the general public.

The best national benchmark available for standards and assessments is the National Assessment of Educational Progress (NAEP) and NAEP subject frameworks (see Appendix B for an example of NAEP's Reading Framework). Based on suggestions from teachers, researchers, subject specialists, administrators, parents and the business community, these frameworks and assessments reflect a general consensus about important competencies and expected levels of performance at key ages and in core subject areas. These frameworks and associated standards are controversial among some policymakers, however, because they were derived from expert opinion rather than judgements of actual student work. Some assessment experts assert that a more effective way to set performance levels for an assessment is to administer the test to students and use actual answers to define performance levels.

Although North Carolina is a notable exception, most state standards and assessments are not aligned with the NAEP purposes, subject frameworks, tests or standards. As a result, states that participate in NAEP may have widely differing results on national and state indicators. For example, whereas 35 percent of Wisconsin fourth grade students were "proficient" on the 1996 NAEP math assessment, 88 percent of fourth grade students were proficient on Wisconsin's own state exam. In Louisiana, 7 percent were "proficient" on NAEP, whereas 80 percent of fourth graders were "proficient" according to the state assessment. Many states reporting results from different programs (i.e., NAEP and their own state assessment programs) run into the problem of seemingly inconsistent results because the two programs have different standards for assigning students to performance levels. Reporting widely divergent indicators of student achievement can be confusing to parents and the general public and send mixed messages to teachers and administrators concerning what content and skills to teach.

Internationally, comparing state standards and assessments to global measures of student achievement is hard work, sometimes requiring

Defining Characteristics of an Assessment

Often, state boards neglect to answer important questions in policy, and responses to these questions are provided later by education agency staff or test contractors. The result may be that the ensuing assessments fail to reflect the thinking of the board. In order to ensure that the state board remains in control of the test agenda, the study group suggests that any state constructing assessments should codify answers to the following questions:

Test Development Issues:

- What is the purpose/goal of the state assessment system?
- Will the tests be criterion- or norm-referenced?
- What content *and* skills will be assessed?
- How many students, and what grades, will be tested?
- What time of year will the tests be administered?
- How many hours should the tests take to complete?
- Are the questions multiple-choice or performance?
- What accommodations for special student populations should be available?
- How will people within the state be utilized in test development?

Scoring Issues:

- What types of scores should be produced and for what purposes?
- What audiences will be primarily using score reports?

Public Engagement:

- How will public opinion be sought?
- Once implemented, how will the public be prepared for the changes implied by assessment results?
- What types of information and training should be offered to educators or the general public in the meaning, purpose and utility of the assessment?
- How will higher education be involved in assessment planning and standard setting?

Program Management:

- What is the schedule for assessment production, piloting and administration?

access to translators and materials from foreign countries. It also requires some knowledge and expertise in comparing foreign education systems. Clearly, most states do not have the resources to undertake comprehensive international reviews of their standards and assessments. It is perhaps surprising, then, that 12 states have looked at student standards, curricular materials, exams or student work in other countries while developing their own standards and assessments (AFT, 1996). For example:

- Delaware, in developing subject frameworks, examined a number of foreign documents, including Australian curriculum frameworks in all core subjects, English science and social studies materials, math materials from The Netherlands, and science documents from Japan.
- In Maryland, the state has arranged for their mathematics and science assessments to be administered to students in Taiwan and Germany so that state assessment officials can analyze differences in the performance of

BEST COPY AVAILABLE

students from the three school systems. The state is also considering translating German exams into English and administering them to Maryland students for comparative purposes.

- Minnesota, Colorado and Illinois participated in the Third International Mathematics and Science Study in order to compare the math and science knowledge and performance of a large sample of their students with students in 41 countries around the world.

While the efforts of these states and others is commendable, most states have had to work with whatever information they could reasonably access, which very often meant comparing their standards with other English-speaking countries that may or may not have high-level standards of their own. The Study Group believes that the field of international comparisons is one which states should not be compelled to navigate alone, since it is unreasonable to expect every state to translate materials and hire international education experts. Instead of the ad-hoc manner in which states have had to measure their standards against those of other nations, the **Study Group calls for a sustainable national effort to provide information and materials concerning educational standards from other countries**. For example, the National Governors' Association and Business Roundtable have developed ACHIEVE, which is a resource center and clearinghouse for governors and business leaders designed to aid them in their efforts to ensure that their standards are "world class."

Finally, some states may choose not to compare their standards and assessments with national or global definitions of "world class." The Study Group affirms a state's right to stand alone. However, even those states that elect not to align their standards with others still need to be able to clearly articulate their reasons for standing alone. Also, these states should be prepared to be called to task by the media and the general public to prove that their standards compare favorably with those of other high quality education systems around the globe.

Conclusion

When state policymakers begin building an effective state assessment system, they are embarking on a complex and time consuming process. This is because, perhaps even more than most other policy issues, the process of building a standards-led education system does not end for policymakers when the ink on the policy dries. **The Study Group emphatically believes that state boards of education need to monitor the implementation of standards and assessment policies to ensure that their policy vision is realized.** This means, minimally, that substantial conversations about standards and assessments should appear with regular frequency on the board agenda. Preferably, these conversations should compare progress in assessment development and implementation with the criteria listed above to ensure that the assessment is as effective as possible. Best of all, policymakers should invest time up-front in clearly defining the type of assessments they envision. What state boards do not define, others will, so staying in control of the assessment agenda means thinking through every important decision about the assessment and defining it in policy. State boards of education that proactively and specifically delineate key characteristics of their state assessments are in the best position to ensure that they drive a system that fosters world class achievement for all students entrusted to their care.

Types of Assessment

Norm-referenced tests (NRTs) compare individual performance against the performance of a representative national sample. Scores from NRTs are usually reported according to "percentile" rank (a student's percentile rank on a test is the percentage of students he or she outscores). This means that students are evaluated in relation to each other rather than to a "standard." Because these tests usually use national averages as a basis for comparison, they are designed to represent curricula nationwide rather than precisely reflect a single course of study. In general, NRTs are predominately multiple-choice; as a consequence they are efficient, economical and require the least amount of subjective involvement by scorers or those setting performance standards. On the other hand, NRTs classify half of students as "below average" rather than giving all students the chance to succeed, usually do not provide practitioners with the data they need to intervene with low performing students, and are unable to measure attainment of content or performance standards.

Criterion-referenced tests (CRTs) compare student performance to clearly defined curricular objectives, standards. Assessment results are usually reported as a pre-defined level of performance or a numerical score. CRTs require the development of meaningful learning objectives that are keyed to assessment items, and CRTs therefore assess what students know and can do rather than how students compare with their peers. Unlike NRTs, it is possible for every student to meet a high standard on a CRT. In general, CRTs are more commonly aligned to state and local standards and curricula than NRTs and may provide more instructionally meaningful reporting of student achievement. CRTs are more expensive to develop than NRTs, because CRTs must be customized to measure specific standards and performance levels.

Multiple-choice assessments require students to select their responses from among a set of specific, pre-determined choices. When carefully designed, multiple-choice assessments can provide reliable information about what students know. Usually, multiple-choice assessments are scored by computers, which provide quick, cost-efficient, accurate and impartial results. However, multiple-choice assessments cannot evaluate students' ability to *describe* what they know, *apply* knowledge to real situations, or *relate* content knowledge from one area to another. Consequently, although multiple-choice assessments yield very reliable scores, in states and districts where learning is not regularly induced through multiple-choice methods, the validity of multiple-choice assessments may be low.

Performance assessment requires students to formulate an original response to a question and to communicate that response through the performance of some act. For example, a performance assessment may ask a student to produce a written essay, model, diagram, or persuasive speech. Performance assessments are usually criterion referenced. Educational reforms and standards that emphasize *application* of knowledge and *expression* of ideas are most effectively evaluated with performance assessments. However, assessment experts continue to debate the overall reliability of various types of performance assessments. Also, performance assessments tend to be time consuming to complete, and so are unable to cover a wide breadth of content. Finally, performance assessments tend to be less efficient and more expensive than multiple-choice assessments because they must usually be scored manually, so creating standardized scoring protocols for performance assessments is a challenge, ensuring consistency among scorers is difficult, and paying for scorers' time is costly.

BEST COPY AVAILABLE

CHAPTER III. Implementation Issues

A. Communications and Assessments: Politics, Public Engagement, Reporting Results

As with the standards on which they are often based, state assessments exist not just in the world of education, but in the public and political realms as well. In some states, assessment systems have been developed in a climate of relative harmony and general public support, while in others new assessment plans have crashed and burned before getting off the ground. In the most notorious case, California's short-lived experiment with large-scale performance assessments (the California Learning Assessment System, or CLAS) was vetoed by Governor Pete Wilson in 1994 just a year after the first round of tests were given, and the wrangling over what will replace CLAS was still continuing as of August 1997. Yet this is not the only recent example of a new state test that became the center of a public and political controversy.

- In 1995 the Arizona Student Assessment Program (ASAP) was suspended by the state superintendent because initial results from the performance assessment part of the program, which asked students to integrate knowledge across subject areas, did not correlate with the results of another more traditional part of the assessment program. Consequently ASAP was substantially changed, and the state legislature later mandated norm-referenced testing in grades 3-12.

- In 1995 the state legislature killed Indiana's newly developed Indiana Performance Assessment for Student Success (IPASS), which included a combination of multiple-choice, short answer, and essay questions, on grounds that the subjective nature of the scoring made it too unreliable and the amount of funding needed for teacher professional development and scoring made it too costly. IPASS was replaced by ISTEP+, a largely multiple-choice test with a smaller performance component.

- Kentucky's ground breaking performance-based assessment program has been the subject of several high-level evaluation panels and also became a campaign issue during the state's 1995 gubernatorial election. Even as the state began distributing millions of dollars to teachers in schools that met improvement targets on the tests, a 1995 panel found that the assessment system was not reliable enough to use for such high-stakes purposes as determining rewards and sanctions for individual schools. Despite continuing adjustments to the program by state officials, a 1997 legislatively-appointed panel, which had to use a professional mediator to help the group deal with sharply-contrasting viewpoints, provided another series of recommendations. These included a proposal that the test concentrate more on basic skills and core academics while not excluding the thinking and writing skills it now highlights.

- The ongoing battle over assessments in California produced hard negotiating between Governor Wilson and legislators over such questions as whether the tests will be only in English; who will oversee the assessments (the governor-appointed state board or the independently-elected state superintendent); and whether the assessments will be off-the-shelf tests or be closely aligned with the standards embodied in California's curriculum frameworks.

Why is there now so much turmoil over an issue that in years past would hardly have caused a ripple in newspapers across the states, let alone create front-page headlines or be the focus of high-level negotiating in legislative chambers and

governors' mansions? The answer is that a number of factors ranging from social and political trends to results of new education research have come together at once to shape this very public debate on what was once a very cut-and-dried issue. Understanding the complexity of these issues can help policymakers and educators better communicate and work with other state leaders and the public on the development and implementation of a new assessment system.

The Issues

1. New high-level standards call for tests that will also be new to parents and politicians.

Underlying much of the turmoil is the fact that the content and the types of items used in new assessments are likely to be new to parents and politicians alike. Because standards in most states are being set at new, higher levels (asking students to demonstrate both a grasp of more knowledge and the use of thinking and communication skills), assessments that match these standards will have items that require students to actively show their ability to think, write, and solve more complex problems. Such open-ended, performance-based items have typically not been seen in large-scale K-12 assessments prior to the 1990s. Even many multiple-choice items will require more complex reasoning than is generally seen in minimum competency or basic skills exams.

Asking students to think and demonstrate their thought processes is not especially controversial in an area like mathematics, where the content is generally abstract. But in subjects like English, where students may be expected to find the "meaning" in a poem, or in history, where students may be asked to write about social trends such as the rise of labor unions, measuring a student's ability to reason and communicate becomes trickier. Some people fear that the line between reasoning and "values" is too thin. Such exercises might be fine for classroom discussions, they say, but when it is raised to the level of a state-sanctioned and scored exam, there is reason to worry about government intrusion into personal matters.

Thus, educators and policymakers must be able to explain why it is important for the new tests to be different from the standardized multiple-choice tests of years past and how this fits into the overall vision for education in the state. True, in many states this same discussion might have taken place during the development of new standards, but the truth is that most citizens were probably not paying attention. When the substance of these standards is embodied in a test, however, especially if there are stakes like graduation involved, more parents and politicians are likely to take notice, and the discussion will have to be repeated.

2. Reporting of test results will also be different from past practices.

Not only is the content and type of some of the test items new to the public, so too is the reporting of test scores. Norm-referenced tests that compare student achievement relative to other students are being replaced by scoring that measures student achievement against a predefined standard. For the state, what is important is having as many students as possible in the "proficient" or "advanced" levels, not where a particular student ranks against all others. The challenge for policymakers is not only ensuring that everyone understands what the results mean, but also assuring those who fear that performance levels will be set too low ("dumbing down" expectations) that the state is truly serious about high-level standards, not just recycling the minimum competency exams created in the 1970s and 1980s.

At the same time, the new, more demanding tests, with "proficiency" set at a higher level, will almost certainly cause overall test scores to be low until students and teachers adjust to the new expectations and become familiar with the format of the tests. Yet some educators fear that such low scores will only further erode the public's confidence in the education system and provide critics with more ammunition to make their case that public schools are failing.

3. It is more difficult to certify the validity and reliability of some types of performance tests for high-stakes purposes.

Despite the potential of performance assessments to inform instruction and more accurately measure the high-level skills demanded by many states' standards, the reality is that the science of large-scale performance testing is still new and rapidly evolving. Thus, while many test developers assert that constructed-response exams and even writing samples can now be scored just as reliably (though not as cheaply) as multiple-choice tests, when it comes to portfolios or other more innovative measures of learning, assuring validity and reliability has been much more difficult to achieve. Only very recently has there been evidence that with careful development of portfolio parameters and evaluation guidelines, as well as extensive teacher training, states can achieve fairly consistent scoring of portfolios on a large scale.

Given the technical challenges, it is not surprising that a range of difficulties have plagued states that were pioneers in performance assessment. For example, in 1995 one of the panels reviewing Kentucky's assessment system found that it was in need of a number of corrective actions and concluded that the state tried to do "too much, too fast" (Harp, 1995). Technical questions also had a hand in sinking CLAS in California, as well as the performance assessments in Arizona and Indiana. While Kentucky officials continue to address concerns and make improvements in their test and the assessment system, there is no question that ongoing difficulties in any state can erode public confidence in an assessment and its uses. When an assessment system begins to be described in the press as "the beleaguered" test (Harp, 1997), it's clear the state faces a communications challenge — no matter what the quality of the test.

4. Competing philosophies of teaching and learning.

The continuing debate over the use of two fundamentally different approaches to teaching and learning has also contributed to the current turmoil over state assessments. Put somewhat simplistically, this debate pits those who favor the teaching of basic skills, fundamental facts, and breadth of knowledge against those who favor allowing young people to construct knowledge from primary materials, thereby providing students with a greater depth of understanding of a smaller set of topics. While this dispute has simmered at the classroom and school level for many years, the current drive toward state standards and assessments has both highlighted the discussion and raised the stakes of the outcome of the debate. Both sides view standards and assessments as an opportunity to embed their philosophy into the state's education goals.

5. Mistrust of the public school system and government in general

Standards-based reform has occurred in an era of growing doubt about the ability of governmental bureaucracies to solve problems, including how to improve education. As discussed earlier, the standards and assessment movement is in part an attempt to address this concern as it moves decision making away from the bureaucracy by giving local schools more flexibility to develop and implement their own strategies — as long as they can meet the state's overall performance goals. Yet ironically, setting the standards and developing assessments is itself a major governmental undertaking that in some cases has only reinforced the public's mistrust.

It is also true that many states and departments of education exacerbated public mistrust when their initial attempts at developing standards and assessments produced outcome statements that were too vague or jargon-laden to be easily understood or used, focused on affective goals as much as academics, or were developed without sufficient public engagement. In fairness it should be noted that standards and assessment development on a large scale is new to most school systems in this country, and many states are gradually refining and improving their work. However, as the experience in California shows, it is not easy to gain the public's confidence after it has been lost.

Lessons Learned

The issues cited above and the rough experiences of some states can be disheartening to policymakers who, after all, are trying to improve the system for all children in the state. The good news, however, is that the experiences of these pioneers in assessments and standards-based reforms have led to thoughtful evaluation by people in and outside of education and generated a number of valuable guidelines that states can use as they continue these reforms. Following are some of the key lessons that have been learned.

#1. Broad public engagement is as necessary in the design and implementation of assessments as it was in the development of standards.

Because state assessment systems based on standards will be just as new to the public as the standards themselves — and because the assessments are likely to have consequences that will affect both elements of the system and individual students — the broad engagement of key constituencies in the assessment development process is imperative for this reform to be credible and successful. These constituencies should include teachers, students, parents, school administrators, school board members, higher education officials and admissions officers, education researchers, state legislators, the governor, representatives of the business community, and the media (CRESST, 1997).

It is important to engage these groups **early** in the process so that state officials can identify their specific concerns and have the groups assist in appropriate ways with development of the overall assessment system and even the tests themselves. As one former state public relations official advised, "Consult *before* deciding," don't just try to inform the public and "spin" the story to attract support after the assessment system is in place (Swift, 1997). Early input by these constituencies will not only bring stronger public support, the policies themselves are likely to be stronger.

#2. Keep the test development process open to public scrutiny.

Secrecy, driven partly by the need to retain the integrity of test items and to avoid the expense of developing new items, "has long been a trademark of the measurement community" (CRESST, 1997). Given the high stakes and new approaches of many of today's assessments, along with underlying skepticism of government reform efforts, policymakers have found the public "increasingly reluctant to accept assessment — new or otherwise — on blind faith." Indeed, critics point out that the secrecy surrounding the development of CLAS items was one of the principal causes of the California assessment's downfall, as rumors spread about "weird things on the test," including content that invaded students' privacy and undermined families' values. By the time the state department released actual items from the test — and many parents conceded the content was not what they had feared — considerable damage to the test's credibility had been done (CRESST, 1997; Kirst & Mazzeo, 1996).

Another problem with keeping the test closely guarded within the confines of a small group or department is that it increases the chances that a set of narrow interests or viewpoints will prevail, for example, either firm believers in authentic assessments on the one hand or "back to basics" advocates on the other. According to one test developer who has worked with a number of states, "Catering to either extreme is a recipe for disaster," and there is a middle ground both in terms of teaching and testing that can be accommodated in an assessment system (Kahl, 1997). An open process can ensure that all sides are heard and a consensus reached.

#3. Take your time.

Many commentators have noted that the development *and refinement* of new standards and assessments will be a long-term process. This is inevitable given that many standards embody a new curricular approach and the accompanying

assessments represent new technology generally untried on a large scale. But because standards and assessments have a strong political dimension — and politicians frequently place policy imperatives ahead of technical feasibility — some states have run up against the consequences of a time line that is too short. As McDonnell writes, "In both the California and Kentucky cases, the assessment designers had to balance daunting technical challenges against the political reality that elected officials expected to see performance assessments on-line within a very short time frame. [At the same time,] there is no question that [education department] officials in both states over promised on what they could deliver. Given the constraints...considerably more time was needed for test development." (McDonnell, 1997)

In addition to technical challenges there are several other reasons for using an extended time frame.

- Training and providing professional development for teachers, who will be expected to understand the new standards and assessments and incorporate new curricular approaches into their classroom strategies, will take considerable time.

- Using an extended time frame and an incremental approach can help the public and politicians adjust to a new assessment system. This can be especially important if the state is having trouble building a consensus or the "political will is lacking to make a needed long-term investment" in new assessments (McDonnell, 1997). An incremental approach could mean that new performance items are only gradually added to state tests (assuring the public that the new curriculum is building on, not bypassing, the basics) or that new tests are given for several years on an experimental or pilot basis with no stakes attached. This would also allow the state to collect baseline data.

- An extended time frame reinforces the reality that new standards and assessments are a fundamental but long-term reform, and that development of new assessments will continually be informed by new research in testing and in teaching and learning, as well as by the successes and problems in other states. As the AFT points out in *Making Standards Matter*,

> All of this will take time. Some states will need to refine their standards and assess several times before they will get them right. That's to be expected. This is a complicated endeavor and it is not reasonable to expect that every state will get it right on the first try. In high-achieving foreign countries, where standards and assessments are at the heart of the education systems, these things have evolved and improved over many years. Ours will need to do the same. We hope the public and policymakers will appreciate this process and show some patience (Gandal, 1995).

#4. Some critics will never be convinced.

Good policymaking calls for boards to hear the concerns of and respond to all groups in the state. However, it is possible that even with the best plan there may still be several vocal but small minorities that will not approve of the policy. If this is the case even after concerted efforts to listen and seriously respond to all opinions, policymakers should not allow this criticism to halt their work. It is a public relations maxim that communications efforts should concentrate on reaching those whose minds are not yet made up rather than those who will never be convinced. As S. Paul Reville, executive director of the Massachusetts Business Alliance for Education and a former state board of education member, writes: given the wide divergence of views on public education issues, over-reliance on public engagement could result in stasis as decision makers try to translate these myriad interests into policy, eventually sapping the popular will for change (Reville, 1996).

#5. There is no substitute for strong political leadership.

Several commentators have noted the importance of strong leadership in building support for

and understanding of a new system of standards and assessments (Reville, 1996; McDonnell, 1997). Given that standards and assessments are a work in progress, leaders will have to exhibit the political courage to champion a plan with a time frame that extends well beyond the next election. They need to communicate to the public that the fundamental ideas of the system are sound, while at the same time remaining flexible enough to work with everyone involved to make adjustments to the policy to reflect new public concerns, new research, or the reality of implementation. Support for politicians taking on this tough role will make ongoing communication among the state board, education department, local school districts and political leaders all the more important.

Reporting Results

One continuing challenge in communicating about assessments is reporting to the public on results from tests *as well as other* indicators of school system performance. Part of the difficulty is that accurate and useful data should include more than just basic test scores, even though most people and especially the media have little patience for detailed analysis and contextual data. Another difficulty is that state reports must be able to speak to multiple audiences (policymakers, educators, parents, the business community), each of which may be interested in different facets of the results or in different levels of detail. An additional issue is the timing of the reports: if tests are reported for individual schools or students, the earlier in the school year the results are received the better, so that teachers and administrators can make midcourse adjustments to curricula, programs and instructional strategies. Depending on the specific purpose of an assessment, however, this may not be possible, such as with an "end of course" exam.

It is clear, then, that successful reporting calls for careful planning that at times will require the state to find a balance among conflicting aims or the needs of different groups. While plans will differ from state to state depending on the components of the particular assessment system, the degree of change from the old system, and the level of public support, the Study Group makes the following general recommendations regarding state assessment reports:

- Results should be reported in straightforward language that is understandable to the broad spectrum of different audiences. Education jargon and technical terms should be

Ten Commandments of Effective Communications for State Boards of Education

1. EXPECT OVERSIGHT/SCRUTINY AS PUBLIC AGENCIES.
2. USE THE HIGH PROFILE OF THE STATE BOARD OF EDUCATION TO PROMOTE YOUR AGENDA.
3. FOCUS ON A FEW KEY GOALS FOR COMMUNICATIONS SUPPORT
4. ENGINEER CONSENT: ALIGN YOUR GOALS WITH THE PUBLIC INTEREST.
5. FOCUS ON KEY PUBLICS ON WHOM THE SUCCESS OF YOUR POLICIES DEPEND. LEARN THEIR CONCERNS. SHARE THEIR GOALS. ADDRESS YOUR JOINT OBJECTIVES.
6. PROMOTE THE VISION: DEVELOP A YEAR-ROUND PUBLIC RELATIONS PROGRAM. INCLUDE LOCAL SCHOOL BOARDS/DISTRICTS AS PARTNERS IN THE COMMUNICATIONS PROCESS.
7. PROVIDE SAMPLE MATERIALS, EXAMPLES OF RESULTS AND CONSEQUENCES. HELP THE PUBLIC SEE WHAT NEW POLICIES WILL MEAN FOR STUDENTS AND FAMILIES.
8. FRAME THE MESSAGE. FIND COMMON GROUND. REPEAT THE THEMES. BE CONSISTENT.
9. TAKE ACTIONS IN ACCORD WITH THE VISION. MANAGE CONTROVERSY. SHOWCASE PROGRAMS. SHARE SUCCESS.
10. ASSESS RESULTS. ADJUST YOUR PLAN. MEASURE SUPPORT. LOOK AHEAD.

BEST COPY AVAILABLE

avoided, and when they must be used these words and phrases should be clearly defined. Language in the reports should be checked for bias.

- State reports should help the public understand the purpose of the data and what the results mean in the context of the overall education system and its goals.
- Participation rates and data from all test takers should be included in state reports.
- Reports on school or district performance (such as school "report cards") should include contextual data and other indicators, not just test results. This could include, for example, percentage of students on free- or reduced-price lunch; indicators of student transience (percent who transfer into or out of the school over a year, or how many remain at the school for three or more years); student to staff ratios or average class size; dropout and attendance rates; and staff experience.
- Reports should be timely so that audiences can see results as relevant and act on them.

B. Including All Students in the Assessment System

Structuring education around outcomes, embodied in standards and measured by assessments, requires policymakers to ensure that all students are learning to high levels. This is particularly true for students with special needs, where there has long been an emphasis on "inputs" rather than "outcomes." **The Study Group believes that states need to include *all* students in assessments to extend to everyone the possibility of receiving diplomas and other life opportunities that test results can provide.** This implies that previously marginalized groups, such as students with disabilities and students with limited English proficiency, need to be included in measures of achievement. Inclusive measures of what students know and can do are important for several reasons:

- To promote high expectations for *all* students;
- To ensure that all students benefit from educational reforms;
- To provide a comprehensive and accurate picture of educational achievement in a state;
- To make accurate comparisons across schools and districts; and
- To avoid retention increases or excess special education referrals that all too often occur before the administration of state tests.

State Assessment Systems and Students with Disabilities

Historically, neither NAEP nor most state assessments have included many students with disabilities. National NAEP exclusion rates of students with disabilities average 50 percent[3], and most states include fewer than 10 percent of students with disabilities in their assessments (McGrew, et al., 1992). Assessment results from the students who do participate are rarely reported on any level.

There are many reasons why students with disabilities have been excluded from state and national assessments. In many cases, guidelines concerning exclusion are vague and a system to monitor exclusion decisions nonexistent. In other cases, accommodations are not made available

[3] Although the national average of exclusion on the 1990 Trial State Assessment of NAEP is 50 percent, individual state exclusion rates range from 33 to 87 percent. A 66 percent exclusion rate was documented for the 1992 Trial State Assessment. Recognizing that the high and varying exclusion rates rendered state comparisons less valid and neglected to measure the achievement of many students, some accommodations were made available, on a trial basis, to a subsample of students in the 1996 NAEP mathematics and science assessments. Preliminary results were promising, and other NAEP assessments may offer some accommodations in future years; however, thorny issues remain, including questions about the validity of results when accommodations are used and about the comparability of results to those obtained under standard conditions.

because of perceived logistical problems and added costs. Some districts discourage participation of students with disabilities because they fear it may result in low scores that reflect badly on some schools. Other districts may purposefully diagnose a student with a disability simply to be able to exclude him or her from the mandatory cohort of tested students.

The result of widespread exclusion of students with disabilities is that few states have a good idea of how effective their special education services are, and even fewer have any meaningful mechanism for holding districts accountable for the outcomes of their special education programs. Simultaneously, countless students leave school without having been given an opportunity to demonstrate what they can do.

Finally, neglecting to test and report test scores of students with disabilities sends a message to students, families, teachers and the general public that students with disabilities are not expected to achieve to high standards.

Indeed, federal law now requires the participation of all students with disabilities in state assessments. The 1997 Reauthorized Individuals with Disabilities Education Act (IDEA) specifies that children with disabilities must be included in state and district-wide assessments, with appropriate accommodations and alternate assessments when necessary to give students a fair opportunity to demonstrate their knowledge and skills. Beginning in 1998, students with disabilities who are able to take assessments with or without accommodations must be tested. In 2000, students requiring alternate assessments must be tested.

This legislation has profound implications for state policymakers, legislators, teachers and students, including issues of cost, reporting, accountability, eligibility and monitoring.

Alternate Assessments and Assessments with Accommodations

Ideally, assessments allow students to demonstrate their knowledge and skills. In order to do so, some students with disabilities may require different test questions (an alternate assessment) or test circumstances (an assessment with accommodations).

Alternate Assessments seek different information and measure different skills than those represented on regular assessments. Alternate assessments are most commonly used with students who are unable to access the standard curriculum, such as students with severe cognitive disabilities.

Assessments with Accommodations are the same assessment as that given to the general student population, but the administration of the assessment is altered in some way. Accommodations are commonly organized into four groups:

- **Format and equipment accommodations** include the use of braille, large print, sign language, communications devices, calculators, word processors, interpreters, pointers and amplifying equipment;
- **Setting and administration accommodations** include settings in hospitals, homes, carrels and small groups and special seat locations;
- **Scheduling accommodations** include giving a test at the best time of day for a student to function, altering timing, stopping testing sessions when a student can no longer sustain activity, giving students extended time and giving the student breaks away from the testing area; and
- **Response accommodations** include alternative methods of responding to testing prompts such as with use of a scribe, machine, lined paper, student dictation sign language or communication device.

The decision to provide special testing circumstances should never be construed as an opportunity to unfairly benefit a few students; rather, alternate and accommodated assessments are intended to "level the playing field" so students with disabilities have a fair opportunity to demonstrate their knowledge and skills without impediments. As such, students should be provided the same accommodations they use during classroom instruction.

BEST COPY AVAILABLE

Cost. Assessments with accommodations can be costly, as they are produced for a small number of students and may require special administrations. Some policymakers fear that creating alternate assessments can be even more costly, although cost estimates of alternate assessment programs are rare so it is difficult to generalize. Recent estimates suggest that Maryland spends approximately $200 per child for start-up costs associated with their alternate assessment program, and Kentucky's ongoing costs to operate their alternate portfolio assessment program that is administered to students who cannot be tested with the regular state assessment are between $75,000 to $90,000 per year. Published, off-the-shelf adaptive behavioral scales can be administered for about $1.50 per student (not including the cost of scoring or reporting). In general, the cost of assessments will probably vary across states, as do assessment costs in general.

Already struggling to fund previously existing special education mandates, state boards and legislatures are now presented with a new and bigger challenge to include all students in their assessment programs. States need to find a suitable balance, and quickly, between investing in the development of alternate and accommodated assessments based on their standards, investing in their general standards and assessment program, and investing in special education programming.

The good news for states is, first, that there is a wide variety of assessment options for students with disabilities that come with a range of associated costs, and, second, that a vast majority of students with disabilities may not need accommodations. Among those who do, many accommodations are simple, straightforward, and relatively inexpensive. Consider these points:

- Twenty-one percent of students with disabilities have speech and language disabilities, and 51 percent are labeled learning disabled. Most of these students can be included in regular assessments with no accommodations.
- The most common accommodation required by students with disabilities is extra time to complete the test, which requires little more investment than the provision of an additional proctor.
- Estimates are that only about 15 percent of students with disabilities will need alternate assessments. This means that among the total student population, between .5 and 2 percent need alternate assessments.
- The percentage of students requiring accommodations is not known, since the extent of

Maryland's Assessment Program for Students with Disabilities

Maryland's assessment program contains a variety of tests and includes *all* students in one of several tests.

The **Maryland Functional Testing Program** is an untimed series of four basic competency tests–multiple-choice in reading, mathematics and citizenship, and a narrative writing assessment. This test is a graduation requirement that must be completed by the end of eighth grade and is taken by all students, including those who require accommodations. Currently, a new graduation test is being developed that will reflect higher standards.

The **Maryland School Performance Assessment Program** (MSPAP), a single test covering mathematics, reading, writing, science, language usage and social studies, is administered in the elementary and middle grades to all students, including those who require accommodations.

Maryland also has a performance assessment for students with severe disabilities, the **Independence Mastery Assessment Program** (IMAP), which has been field tested in about one-third of all districts in the state. This test reflects a set of outcomes that are life-skills oriented. Students with severe disabilities can participate in MSPAP with no accommodations, MSPAP with accommodations, or IMAP.

their use is dependent upon policies that determine which accommodations are allowed. States are just now beginning to collect information on how many students use accommodations and what the nature of those accommodations are (Ysseldyke, *et al.*, 1994).

There are also several policy avenues that states can pursue in order to decrease expenditures on alternate and accommodated assessments. Most importantly, **states need to devise mechanisms to conscientiously limit the use of alternate and accommodated assessments to include only those students whose knowledge and skills cannot be fairly assessed by standard means.** Measures to ensure that alternatives to regular assessments will not be overused include:

- Removing incentives for districts to label an excessive number of students as having disabilities, such as allowing exemptions to the reporting of test results;
- Restricting each student's allowable accommodations to those that the student uses regularly for classroom instruction;
- Devising specific criteria that limit the potential overuse of alternate assessments. For example, in Kentucky and Maryland, students are eligible for an alternate assessment only if they are working on a curriculum that does not lead to the award of a diploma; and
- Auditing districts when the rate of exemption from testing seems inappropriately high.

The special education community has long called for students with disabilities to be included in assessments that offer the most fair, and least intrusive, accommodations. States can contain costs by making sure this vision is realized and that districts have no advantage in prescribing unnecessary and overly intrusive accommodations.

Reporting. IDEA requires that alternate and accommodated assessment results be reported to the public with the same frequency and in the same detail as the assessment results of nondisabled students. But clear, meaningful reporting of results for standard, alternate and accommodated assessments is a challenge. Currently, of the 48 states that produce statewide education accountability reports, only five produce separate performance reports on students with disabilities. Among the rest, 12 states include separate performance data on students with disabilities, eight states specifically exclude performance data on students with disabilities, and all states have at least one report in which they do not specify whether certain data include or exclude students with disabilities (Thurlow, et al., 1997).

States need to ensure that fair and accurate reporting structures are in place that: 1) neither benefit nor disadvantage schools or districts with disproportionate numbers of students with disabilities; and 2) do not reveal information about individual student scores even though accommodated or alternate assessments may have been given to very few students. Implementing these two conditions can be messy, however: the first requirement can be achieved by separating school performance data for students with and without disabilities; the second requirement, however, will require in some instances (where schools have few students with disabilities) that disaggregated scores for students with disabilities not be reported. States need to develop reporting criteria that balance these two needs.

Accountability. Systems of rewards and sanctions based on student performance have often been the reason that schools and districts excluded students with disabilities from assessments. For the first time ever, however, IDEA regulations will require districts to make available comprehensive data to discern how well schools, districts and states serve populations with disabilities. State boards of education would be remiss if they did not take this opportunity to expand definitions of school and district accountability to include these measures of achievement and create a new system of rewards and sanctions that does not

benefit certain schools with more or fewer students with disabilities. A system that does not factor in the work schools do with students with disabilities is incomplete and inequitable; and a system that is not refined enough can mask pockets of poor student performance within schools.

Eligibility and Monitored Participation. Decisions regarding who takes alternate or accommodated assessments vary widely across schools, districts and states and often vary according to whether the test is "high stakes" or "low stakes," whether the student is likely to perform well on a particular assessment, and whether test scores for students with disabilities are publicly reported. Although decisions about student eligibility for alternate or accommodated assessments should be made largely by teachers who work closely with the student and other members of the student's Individualized Educational Program (IEP) team, states need to provide stringent and comprehensive procedures to guide these decisions and ensure that large numbers of students are not shifted out of regular assessments unnecessarily. States also need to require districts to justify administering an alternate or accommodated assessment and develop a mechanism to monitor and evaluate their decisions to ensure that regulations are being implemented equitably statewide.

Assessment and Students with Limited English Proficiency

America is diverse and growing more so. As of 1990, 14 percent of the school-aged population of the United States spoke a language other than English as their first language, and this percentage is likely to grow through the beginning of the next century. Exempting students with limited English proficiency (LEP) from taking statewide assessments means, in some schools, that fewer than half of the student body is being held accountable for achieving state standards.

LEP Students and Assessments in Federal Law. In 1994, the U.S. General Accounting Office issued a report which had as its premise that "the nation's ability to achieve the national education goals is increasingly dependent on its ability to educate LEP students." Federal guidelines for Title I (programs for disadvantaged students) and Title VII (bilingual programs), which together serve most LEP students nationwide, explicitly state that students served by related programs should meet state and local performance standards. In the case of Title I, the academic progress of students served by related programs *must* be measured against the same standards, and through the same tests, as regular students. Additionally, LEP students must be tested in the language and form most likely to yield valid and reliable measures of their achievement. In essence, state boards and other state policymakers bear legal responsibility for establishing valid and reliable measures of the academic progress of LEP students.

Still, LEP students are rarely included in state assessments. In 1996, only five states required English language learners to take statewide assessments that were required of other students. In three of these five, these students could still be exempted under certain conditions. In 36 states, students with limited English proficiency were exempt from state assessments. In 22 of these states, students were required to take the assessments after a period of time in which it is presumed they learn to speak English, usually one to three years. Several of the remaining 14 states based decisions about the participation of English language learners on the English proficiency level of the student, regardless of years in school. Other states allowed local districts to decide which students have sufficient English to participate in statewide assessments (August and Lara, 1996). In general, there remains no consensus across the states concerning the inclusion of LEP students in state assessments.

Accommodations for LEP Students. At whatever point states do include LEP students in statewide assessments, they differ widely in the testing accommodations they allow. About half of all states allow some type of modification in the administration of statewide assessments to incor-

porate LEP students, including extra time to complete the test (20 states), grammatical and vocabulary simplification of test directions (14 states), small group administration (18 states), use of dictionaries (13 states), and the ability to read questions aloud (12 states) (Hafner, 1995). Ten states also provide guidance to performance test scorers on how to evaluate the written work of students with limited English proficiency. To date, no research has investigated the validity or reliability of these accommodations, so it is impossible to know if the results are comparable to those of other students across the state.

In efforts to give LEP students the opportunity to demonstrate their knowledge and skills most effectively, some states and districts have experimented with alternate assessments, including:

- **Administering nonverbal tests**, which ask students to demonstrate a skill rather than write answers. Unfortunately, student performance depends upon understanding initial instructions and the ability to ask clarifying questions. Consequently, many suspect that scores on nonverbal tests are still affected by language comprehension skills.

- **Developing native language assessments**, which are translations of the original English test into LEP students' first languages. However, performance on these assessments can be affected by myriad circumstances, including regional and dialectical differences within languages, the difficulty of directly translating some vocabulary from English into other languages; and lack of literacy development in the first languages among many LEP students.

- **Modifying question language**, which entails reducing nonessential details in test questions or simplifying grammar. To date, one study has investigated the effect of this type of accommodation on the National Assessment of Educational Progress for LEP student performance. The results were modest and statistically unreliable.

- **Providing Interpreters**, where speakers of the student's native language translate directions and test vocabulary for LEP students. At present, research on the consequences of providing interpreters is negligible and the effect on validity and reliability is unknown.

Title I Students and State Assessment Systems

When Congress passed the *Improving America's Schools Act in 1994*, they revamped Title I and changed the requirements for assessing the progress of students participating in the program. In the past states were required to use a norm-referenced standardized test to measure the performance of Title I students. Now, however, states must use the same assessments to measure the performance of all students in the state and must develop standards and accountability requirements to help all students reach, or demonstrate sufficient progress toward, proficiency. Recognizing that Title I students often do not reach the same standards as other students, states now must define "adequate yearly progress" among students served by Title I funds in terms of their performance on state standards-based assessments, determine whether individual schools are making adequate progress with Title I students, and target an appropriate date by which all students served by Title I will perform at a proficient or advanced level as defined by state performance standards. Based on this target date, states must set annual rates of improvement on test scores to achieve their goals.

To encourage accountability for students who are most at risk, the law requires that states disaggregate student scores in districts and schools according to gender, race, ethnicity, English proficiency, and migrant status and enable comparisons between nondisabled and disabled students and economically disadvantaged students and advantaged students.

BEST COPY AVAILABLE

The Study Group agrees that, fundamentally, **every assessment is a language assessment**. This is even more the case as states develop performance assessments that require high-level language skills. In general, there are impregnable challenges to accurately measuring the achievement of LEP students, much less to do it in a way that measures their performance against content and performance standards.

Schools have long struggled with ways to disaggregate the complex interaction between language proficiency and subject knowledge among LEP students. As states develop assessment and accountability systems, they too are compelled to contend with a dearth of knowledge concerning the most effective way to include LEP students in state assessment programs. From a policy perspective, this task is filled with challenges, including:

- **Spearheading public conversation** and encouraging consensus-building about fair and informative ways to assess the knowledge of LEP students;

- **Coming to agreement, both within and among states, on the definition and classification of LEP students**. Currently, definitions of "English proficiency" tend to be unregulated and unstandardized across classes, schools, or districts;

- **Devising a fair method to assess LEP students**, in light of the lack of legal and experiential documentation of effective methods to include LEP students in assessments;

- **Overcoming the heretofore inconsistent or nonexistent record-keeping of assessment participation rates among LEP students**;

- **Standardizing local decisions** about allowable assessment accommodations;

- **Eradicating the widespread habit of deferring assessment** until students have a high level of proficiency in English. Neglecting to test students until they are proficient in English denies states and districts valuable data about the progress of LEP students and fails to hold schools accountable for student progress occurring within bilingual and English as a second language programs;

- **Establishing interim reference points.** Depending on their age upon arrival in the U.S., it may take five or more years for a student to begin to achieve comparably to native English speaking classmates. (Collier, 1987; Cummins, 1981). Especially in the first four years, academic progress is slowed by the formidable task of learning English. States that have widely disbursed benchmarks may find that the more subtle progress of LEP students is not reflected by their measuring system. Establishing smaller, interim performance indicators within widely disbursed benchmarks may measure LEP students' gains in knowledge more accurately.

- **Training school staff and assessment scorers in LEP expression patterns**. Educational staff and assessment scorers sometimes evaluate students' English expression. When this happens, students are evaluated according to their language skills rather than subject knowledge. States may want to consider investing time and funds in training all professionals involved in student evaluation to take into account the English patterns of emerging English speakers.

- **Disaggregating scores.** To determine whether LEP students are meeting standards, and to hold schools accountable for "adequate yearly progress" for all students, states need to separate scores according to English proficiency status. Currently, Florida, Hawaii, Louisiana, Maine, Ohio and Washington do this.

- **Developing standards in English as a second language.** Establishing standards in English as a second language provides schools, districts, and the state a gauge to measure the progress of students' English language acquisition. It

also provides a tool to measure the extent to which schools and districts help students develop fluency in English. States that take seriously the task of establishing high standards for all students should consider developing standards in English-language arts for LEP students.

Acknowledging that there are substantial obstacles to including LEP students in statewide assessments, and in concordance with a recurring theme in this report of including all students in state assessments, **the Study Group asserts that state boards need to provide leadership in proactively developing policies to include students with limited English proficiency in state assessment systems.** In this instance, the Study Group's recommendation requires state boards to extend their policymaking into an area where scant evidence provides examples of "good practice."

A word of caution, however: being in the vanguard means taking risks and expecting mistakes. It means that state boards need to be able to patiently persevere in making long-term and incremental change as the research and policy world learns more about effective LEP assessment. It also means that state boards need to withhold imposing sanctions or reactionary policies according to unclear and immeasurable assessment results among LEP students. In addition, it means that states should consult with bilingual community groups, teachers, administrators and others to devise a clear and consistent assessment policy that is as fair and informative as possible while recognizing that assessing LEP students will, for at least the near future, be a work in progress. Most importantly, it means being able to engage in dialogue, as a board and with the public, about what the results mean and what the results are measuring. Finally, it means welcoming researchers to investigate the effectiveness of vanguard assessment policies in improving the quality of assessment results for LEP students.

Federal law is moving states rapidly toward an accountability system that includes all students. But this is a sea change from even five years ago, when educators were welcome to hold different expectations for different students and when LEP and special education students were taught and evaluated in ways very different from other students. State policymakers need to act quickly to ensure they have a data collection, reporting and intervention system in place that is fair to students, teachers, schools and districts.

C. Professional Education

"States need to provide time and money to re-educate teachers and administrators for good use of assessment results. Make front line people assessment literate so that assessments can be used to improve what happens on the front line. In other words, I ask that you don't think of state assessments as an opportunity to sanction front line people, but an opportunity to educate them."

— Ron Areglado, National Association of Elementary School Principals

As is true with any education reform that is attempting to impact instruction in the classroom, the development and implementation of a state wide assessment system will not succeed without the involvement of teachers. Engaging educators from the beginning of the process not only promotes "buy-in," it will likely result in a better, more stable system. In addition, if specific purposes of the assessment system include improving instruction and helping students meet the higher standards, then again success will depend on teachers' understanding of the content and instructional implications of the standards, their understanding of how to shape their classroom assessments to align with the standards, and their understanding of how to use the results of state assessments to adjust their instruction for individuals and for their students as a whole. At the same time, if there are consequences for schools or districts attached to the assessments, it is only fair for all school staff to be fully informed about the assessments and represented in the development of the system.

Finally, teachers remain parents' strongest link to the education system, as well as their primary source of information about their child's achievement (ECS, 1997). This means that individual teachers' understanding of state assessments is critical both in terms of providing accurate and helpful information to parents about their child and in communicating the state's overall education goals to parents and the public. It would be foolhardy for the state to ignore this means of connecting with its most important constituents.

Following are some of the specific ways states should involve educators in the planning and implementation of the assessment system and providing professional development in this area:

- Include educators on committees that set performance standards;
- Include representatives of teacher and administrator organizations on state panels or task forces involved in planning the assessment system;
- Sponsor workshops at the school, district, and state levels to inform educators about the assessment system and provide instruction in general assessment theory and practice (especially in performance assessment). Workshops should include specific guidance and sample test items to enable teachers to translate state standards and curriculum frameworks into instructional strategies, course content, and classroom assessments.
- Involve teachers directly in developing and piloting performance assessment tasks and in scoring performance-based sections of an assessment. In Maryland, nearly half of all fifth-grade teachers and 70 percent of eighth-grade mathematics teachers participated directly in some aspect of developing, piloting or scoring the Maryland School Performance Assessment Program (MSPAP). This type of professional development was more frequently rated by the teachers as being "very useful" than were the informational sessions (Koretz, et. al., 1996). In fact, many teachers say that workshops focused on scoring performance assessments are the most valuable professional development experiences they have ever had (Rothman, 1995).
- Sponsoring workshops at all levels regarding changes in instruction and curricula inherent in the new standards and assessments;
- Distributing print and video resources related directly to the assessment as well as to changes in instruction and curricula; materials may be developed by the state, districts, or by regional or national professional organizations;

Professional Development: It's Not Just for Teachers

Standards and assessments are at the core of the state's education system. This means that local school board members, district personnel — as well as members of the state board of education — must fully understand the state assessment system. State and regional workshops are a key element of this effort, and they should always include "hands-on" experiences with actual test items and examples of student responses along with scoring guidelines, or rubrics.

As with inservice training for teachers, this professional development will be more effective if it is developed as a series of workshops rather than a one-time event. This is particularly true with training for a complex assessment system because it has many components, it will be rolled out over a number of years, and it is likely to evolve over time. Indeed, in future years such workshops could be structured for two-way communication so that district and school-level policymakers and administrators could provide feedback to the state on problems, successes and suggestions for improving the system.

BEST COPY AVAILABLE

- Encouraging teachers to meet and work collaboratively on planning new instructional strategies related to the standards and assessments or in preparing for the assessment.

Assessment and Preservice Training

> *"Teachers need deep knowledge about assessment. Now, most teachers can't tell you if a test is good or not."*
>
> — Joan Baratz-Snowden, American Federation of Teachers

> *"Most states have no assessment knowledge requirements for incoming teachers, and in particular they have no requirements for them to become competent in performance and classroom assessment."*
>
> — *Testing Our Children: A Report Card on State Assessment Systems* (FairTest, 1997)

One other area of teacher training regarding assessment also needs attention by state boards of education: instruction in teacher education programs. Study Group members heard time and again from expert witnesses that teachers first entering the classroom were critically short of knowledge about assessment, whether it was assessment theory, how to develop and use assessments in the classroom to improve instruction, how to analyze and use results from large-scale assessments, or how to develop and score newer forms of performance-based assessment. Such deeper knowledge about testing is key if educators at all levels are truly going to use assessment to improve learning and instruction, rather than just to assign a grade or rank students and schools.

State boards of education have several policy mechanisms available to ensure that knowledge of assessment and skills in developing and using tests are part of every beginning teacher's foundation. These include:

- **State standards for teacher licensure**: New standards for teacher licensure now being developed by many states provide state boards with a powerful tool for improving beginning teachers' familiarity with assessment. Questions state boards should ask include: Are standards for teacher licensure aligned with academic standards for students? Is understanding and use of student assessments included as a component of the knowledge and skills necessary for licensure? Do the assessments leading to licensure mirror assessment practices states are promoting for use with K-12 students (i.e., multiple measures that include observations, performance tasks, basic skills items)?

- **Partnerships with higher education boards and other organizations involved in teacher preparation**: State boards can use their leadership position to work with state and national groups involved with teacher preparation in promoting a standards-based licensure system that requires knowledge of student assessment. Such partnerships can also affect the program accreditation process.

- **Other state structures supporting teacher development**: State boards can promote knowledge of academic standards and assessment practices through other state structures that support teacher interns or beginning teachers (such as mentor or master teacher programs).

Cost is yet another issue state policymakers need to consider regarding professional development, both in terms of time and money. Given the importance of standards and assessments to the education system overall, the **Study Group recommends that states assume the burden of paying for initial training regarding the new assessment system**. This should include covering the costs of teachers and administrators involved in the various stages of assessment development; producing and distributing printed materials and videos about the assessments; sponsoring regional workshops about the assessments for teachers,

administrators, school board members and district personnel; and helping schools pay for release time for teachers attending workshops and other inservice sessions regarding the assessments.

D. Linkages with Higher Education

While state policymakers are moving to an outcomes-based system of standards, assessments and accountability measures, most colleges and universities continue to make admissions decisions based upon grade point averages, class standing, Carnegie units, required courses and standardized, norm-referenced test scores. Why have colleges and universities failed to adopt standards-based admissions, in line with the standards-based assessment and accountability systems being developed by states? According to David Conley of the University of Oregon and Oregon State System of Higher Education, universities have neglected to rally around the changes taking place in states for several reasons, including:

1. Higher education has always drawn a distinction between conceptual and applied knowledge; universities have historically preferred the former, and school reform seems more concerned with the latter;

2. Admissions officers resist anything that may complicate the admissions process, and they resist portfolios in particular; and

3. Higher education does not have very high expectations of the secondary school system in general and believes that the calls for "systemic reform" are a smokescreen for lowering standards.

With hundreds of thousands of high school graduates every year potentially affected by the acute disjuncture between emerging expectations of high school graduates and entrance requirements of colleges, it is more important than ever that the transition between the two levels be systematized and coordinated. Yet in many states, colleges and universities have been noticeably absent during the standard setting process. In other states, colleges and universities have contributed to the standard setting debate but still require incoming students to take a battery of placement exams rather than listening to what state assessments say about a student's ability to do college level work.

The Study Group believes it is time for colleges and universities to be centrally involved in the standard setting process, and in assessment production. College admissions should be systematically addressed in state board discussions of standards and assessments. In return for a key place at the policymaking table, **the Study Group recommends that colleges develop systems to accept students based on proficiencies, rather than exclusively on class rank, course work, credits or grade point average.**

A small number of examples of cross level cooperation and collaboration for a standards-based university admissions system exist.

- In Oregon, several public colleges and universities have agreed to admit students based on demonstrated proficiencies rather than on traditional measures. The Proficiency-based Admission Standards System (PASS) provides consistent standards statewide that eliminate grade inflation and teacher subjectivity by scoring student work against common criteria that have been agreed upon by K-12 and higher education representatives.

- In Wisconsin, a pilot program admits students to higher education based on "competencies" such as "use of variables in mathematics," "science processes," and "integrative applications of social sciences." This program was developed largely by higher education faculty members but was reviewed by high school teachers before being implemented. The competency system allows schools to offer instruction that does not conform to subject-area requirements. In place of traditional class grades, teachers rate students on skills that often cut across classes.

- Georgia's P-16 Initiative is a comprehensive, collaborative effort to raise expectations and ensure student success from preschool to higher education. Currently, at least fourteen local councils have been developed for the job of brainstorming how to improve student achievement, facilitate students moving smoothly across educational levels, and assure that students entering higher education have the knowledge and skills necessary to succeed.

State policymakers know too much about what constitutes a "good student outcome" from the K-12 system to allow higher education to ignore these meaningful measures in favor of outmoded and inaccurate indicators of student achievement. They owe it to the students entrusted to them to work in concert with higher education to create a seamless transition between K-12 and higher education.

E. Evaluating Assessment

Assessment and accountability systems are costly — not only in financial terms but also in terms of educators' efforts, instructional time, and future careers of students, administrators and teachers. If policymakers expect the public to support these costs, they must ensure that the assessment and accountability system is effective, efficient and equitable. This means that state boards need to implement responsive program planning, pay attention to the implementation process, ensure that individuals and organizations have the necessary materials to carry out their particular missions and check to ensure that program intentions are realized in program outcomes. Answering these questions requires a mechanism for program evaluation.

Unfortunately, the norm today is that systematic evaluation of a state's assessment systems rarely takes place unless there is overwhelming public or political pressure to do so. The Study Group believes that this is a mistake that can cost states dearly in terms of accuracy, usefulness of their data, efficiency of their testing program, fairness to students and schools, and credibility and public support. The Study Group believes strongly that state boards of education need to hold *themselves* and the state education department accountable for both understanding the effects of an assessment system and ensuring its success. **Because of the power of evaluation to improve policy, the Study Group asserts that state policymakers need to consider program evaluation alongside policy development.** Doing so requires attention to several factors:

Participation. State boards need to give special attention to ensure that those centrally involved in the administration or impacted by the results of an assessment are invited to participate in the design of the program evaluation and in the evaluation itself. Evaluations that collect data from a number of individuals involved in all aspects of program planning and implementation will be more valid and reliable and provide more fertile ideas for program improvement than those that involve only a few constituent groups.

Internal vs External Evaluators. Deciding whether to have department staff conduct an evaluation or contract an external evaluator has many implications for the evaluation itself. Assessments undertaken by department staff are usually cheaper and faster, since staff know the nuances of the program better than someone from outside. External evaluators, on the other hand, may be more objective and removed from any politics surrounding the assessment system. For example, the RAND Institute on Education and Training has conducted evaluations of Kentucky, Vermont and Maryland.

Communications. Effective program evaluation provides information that can be shared with the public to encourage dialogue. On the other hand, when an evaluation points out a need for program improvement, a wide range of special interest groups may co-opt the results for their agendas. To balance this, the state board should ensure that a communications plan is in place to help guide public conversation in constructive ways.

Using results. Before undertaking program evaluation, policymakers must be confident that they have the money, resources and momentum at their disposal to respond to the results effectively and make necessary program changes. Authorizing an evaluation and then neglecting to respond to its conclusions is not only a waste of time, money and resources, but it can jeopardize the public credibility of both the assessment system and the evaluation.

But agreeing to use results can have profound implications on the implementation of a state's assessment system. For example, in a validity and reliability study of their new testing program, Vermont learned that their portfolio assessments, while powerfully and positively affecting classroom instruction, yielded results so unreliable that the data could not be used for most of their intended uses. Based on the evaluation results, the state had to scale back its reporting plans,

Types of Evaluation for State Assessment Programs

Five types of evaluation can inform assessment program improvement at different stages of the process: development, design, process, management and impact.

Development. Development evaluation is undertaken before any assessments are designed. It consists of doing a needs assessment, reviewing practice, and reviewing research to evaluate the problems an assessment program should solve. For example, undertaking development evaluation of student achievement in your state, you may learn that students are not achieving because most schools do not enforce attendance laws, because teaching practice is abysmal or because textbooks are woefully outdated. Comparing your state to "best practice" may indicate that simply developing standards and assessments is not sufficient to improve performance.

Design. Design evaluation judges program design, usually before implementation, and clarifies program logic and feasibility. It addresses questions such as: Do the parts of your assessment program fit together? Are you assessing the right people at the right time and on the right information to reach your stated goals?

Process. Process evaluation studies the implementation of the assessment. It includes such activities as asking teachers, administrators, parents and students what their experience of the assessment was like. Process evaluation illuminates unintended consequences of sc ne tests such as discouraging certain teaching practices, costing districts an inordinate amount in staff and instructional time, or making students too tired to perform well. Process evaluation addresses questions such as: Is the test exceptionally hard to administer? Is it being implemented as designed? Does the test take too much time away from instruction? Do parents feel their children have been unfairly judged? Is the curriculum narrowing or broadening?

Management. Management Evaluation looks for efficiencies and inefficiencies in how the assessment program is managed. It answers questions such as: Is SEA staffing sufficient to carry out this project efficiently? Are human relationships positive among staff? Is the relationship between the SEA and testing contractors as productive as possible?

Impact. Impact evaluation takes place after an assessment program has been operating for several years. In general, impact evaluation examines if the program is meeting its objectives and answers the question: So what? Results from an impact evaluation can be used for accountability, justification and program improvement.

BEST COPY AVAILABLE

only releasing statewide total averages until issues of test reliability could be satisfactorily resolved. (Koretz, et al., 1994)

Comprehensive evaluations of state assessments and accountability systems are rare, but the new emphasis on outcomes accountability for students, teachers, schools and districts requires a similar shift in emphasis to outcomes accountability on a state level. **The Study Group believes that it is time for a sea change in perception among policymakers — assessments need to be evaluated not according to the results they produce, but according to their ability to fuel an accountability system that genuinely improves teaching and learning and encourages everyone to achieve high standards.** Fleshing out these more subtle consequences of an assessment system requires targeted investigation–a mechanism for program evaluation.

F. The Financial Costs of State Assessments

The type of systemic assessment program recommended by the Study Group does not come at small expense. (The accompanying box provides a range of the costs associated with various state assessments). In addition to the tests, there are the costs related to professional education, alternative assessments, communicating with the public, and evaluation. Still, given the importance of an assessment system in the overall scheme of the state's role in education and given the harm that an inexpensive system can do in terms of narrowing the curriculum, diminishing the viability of state standards, or furthering rather than reducing inequities, **the Study Group believes that funding must be sufficient to enable the state to realize the stated purposes of the assessment system.** And such funding should also be placed in perspective: although dollar amounts appear high, for each of the examples in the box at right, funding for state assessments range from about .04 percent to .9 percent of each state's share of K-12 education dollars.

Large-Scale Assessments: What Do They Cost?

Developing an assessment system aligned with state standards requires many types of investment from policymakers, not the least of which is financial. Test developers estimate the cost of administering an off-the-shelf, norm-referenced multiple-choice test to be approximately $5 to $8 per student. But creating an assessment that is aligned with state standards can easily double the cost. Adding performance items to the test brings the cost to $35 or more per student. In general, the cost of state assessments varies substantially depending on the subjects tested, types of questions asked, and number of students tested. For example:

- In 1997, Colorado legislators cut back a $1.6 million first-year price tag for a new K-12 testing program in a system that has 641,000 students. As a compromise, the education department dedicated money from its own appropriation to augment the reduced legislative funding, and the testing program was pared to include only reading and writing. The test has both multiple-choice and performance questions.

- In 1998, Florida's students in grades 4, 5, 8, and 10 will take tests that yield individual student scores in reading and math for school and district accountability. With a K-12 enrollment of 2.2 million students, this program costs approximately $8 million annually.

- In Indiana, the Statewide Testing for Educational Progress (ISTEP+) is a largely multiple-choice test combined with a small number of questions requiring written answers, and it is aligned with Indiana's standards. Given in the third, sixth and tenth grades, this assessment costs $24.8 million, for a K-12 student population of almost 970,000.

- Finally, in Iowa, where an off-the-shelf, norm-referenced test is administered, the number of grades taking the Iowa Test of Basic Skills tripled in 1995, from three to nine. The cost to the state, which serves about 500,000 students, was a comparatively small $500,000.

BEST COPY AVAILABLE

At the same time, state boards must be realistic and pragmatic, understanding that legislatures may balk at the costs associated with creating such a system. State boards may also find that within their own budget considerations there are other priorities that compete with the cost of creating and administering a comprehensive state assessment program.

Within the constraints of each state's budget and political will, there are several ways to design a relatively affordable assessment system that is still effective and comprehensive. First, revisiting a recurring theme of this report, state assessment systems are "works in progress." As such, building the system can be undertaken incrementally so the cost of developing assessments is distributed over several years. Building assessments incrementally is also a good idea because it allows time for incremental change in teaching and learning and steady evaluation and refinement of assessment instruments that may not be feasible in a quickly constructed and immediately implemented assessment system. Other ways in which states can limit the costs of assessments include:

- Using assessments for school and district accountability that are given to *samples* of students. For example, Maryland's performance assessment used for improving instruction and for school accountability (MSPAP), is given to a sample of 3rd, 5th, and 8th grade students, and each student takes only a portion of the exam.

- Considering, particularly in the younger years, using an off-the-shelf skills test to assure parents that their children are developing fundamental abilities in reading and mathematics. However, the Study Group cautions against using such tests for high stakes purposes or even creating "school report cards," as this could too easily lead to narrowing the curriculum and drilling to the test. As indicated above, more expensive tests that use a higher proportion of performance items and are more closely aligned with state standards can be used for accountability purposes (as in the Maryland example) and/or with older students; and

- Developing and scoring state performance assessments locally. By allowing districts to develop (using state guidelines), score and utilize performance assessment scores for program improvement purposes and/or assisting individual students, state policymakers can cut costs considerably and overcome the difficulties inherent in trying to standardize performance assessment scores at the state level.

G. Conclusion

State boards of education have enormous responsibility and authority both in providing the policy framework for standards and assessments and in overseeing the development, implementation and evaluation of state assessment systems. Indeed, the Study Group maintains that this effort provides much of the foundation of the state's role in education and it is pivotal to the state board's leadership position in the system. Thus, the Study Group urges state boards of education to think systemically in their discussions and policymaking around testing issues. State content standards, curriculum frameworks, teacher licensure, professional development, public relations, accountability and evaluation programs should all be closely linked with state assessments.

Finally, it is the Study Group's hope that this report provides state board members with the information and impetus they need to be active, positive participants in ongoing discussions of new assessment systems. Even in states with the most developed structures, assessment systems are generally in a state of flux, constantly undergoing changes ranging from fine tuning to complete overhauls. Given this continuing evolution and the fact that standards and assessments are so central to the state's role in improving education, it is incumbent on every state board to remain involved and assert their leadership in shaping these critical policies.

APPENDIX A. Glossary of Assessment Terms

Accommodations: Alterations made in the administration, not the content, of an assessment in order to "level the playing field" for a student with a disability. Accommodations can include providing a student with special equipment (such as an amplifier), producing the assessment in a different format (such as braille), administering the assessment in a different setting (such as a hospital), altering the test schedule (such as giving extended time), or providing alternate ways to respond (such as the use of sign language).

Alignment: Ensuring that the knowledge and skills asked for on assessments is the same knowledge and skills specified in content standards.

Alternate Assessment: Seek different information by measuring different skills than a regular assessment. Alternate assessments are commonly administered to students with disabilities who are unable to access a standard curriculum, such as students with severe cognitive disabilities.

Benchmark: A subcomponent of a standard that more specifically defines what students should know and be able to do at various grade levels.

Content Standard: What students should know or be able to do in a specific subject area. *Benchmarks* define the point in a student's career (usually a grade level) when he or she should have acquired such knowledge or skill.

Criterion-referenced test: See "Types of Assessment" on page 17.

National Assessment of Educational Progress (NAEP): Also called "The Nation's Report Card," was created by Congress and is administered by the National Center for Education Statistics, U.S. Department of Education. This test is the only ongoing nationally representative measure of what students in the United States know and can do in 10 subjects and three grade levels.

Norm-referenced test: See "Types of Assessment" on page 17.

Performance standard: Defines how good is good enough in terms of achieving a content standard. Frequently divided into levels such as "advanced," "proficient," "basic," and "below basic."

Performance assessment: See "Types of Assessment" on page 17.

Portfolios: A type of performance assessment. Portfolios are collections of student work that demonstrate progress over time and level of performance.

Reliability: Concerns the extent to which a student would receive a similar results if the test were administered again. Reliability is discussed in terms of "consistency" (e.g., will a student's score on a test taken one day be close to his or her score on the test if it is taken the next day) and "generalizability" (e.g., does a student's score on a 60 minute math test accurately reflect his or her knowledge of the entire subject).

Rubric: Scoring guides used for performance items that describe the characteristics or components of responses to questions earning various point values. For example, a rubric would describe what an answer should contain to earn the student a "basic," "proficient" or "advanced" rating on a particular question.

Sampling: Giving a test to a randomly chosen subset of students in order to provide a statistically equivalent picture of how all students in that group would perform. On some assessments, each randomly chosen student takes only a portion of the test. This technique, called "matrix-sampling," is an efficient way to provide reliable group (e.g., school) results.

Validity: The degree to which a test actually measures what it purports to measure (e.g., the extent to which an I.Q. test actually measures intelligence); also, the degree to which particular uses and interpretations of assessment results are justified (e.g., the extent to which a certain score on an SAT or ACT test predicts a student's chances of succeeding in college).

APPENDIX B: Reading Framework from the National Assessment of Educational Progress (NAEP)

1992 NAEP Framework—Aspects of Reading Literacy				
Constructing, Extending and Examining Meaning				
	Initial Understanding Requires the reader to provide an initial impression or unreflected understanding of what was read.	**Developing an Interpretation** Requires the reader to go beyond the initial impression to develop a more complete understanding of what was read.	**Personal Reflection and Response** Requires the reader to connect knowledge from text with personal background knowledge.	**Demonstrating a Critical Stance** Requires the reader to stand apart from the text and consider it.
Reading for Literary Experience	What is the story/plot about? How would you describe the main character?	How did the plot develop? How did this character change from the beginning to the end of the story?	How did this character change your idea of X? Is this story similar to or different from your own experiences?	Rewrite this story with _____ as a setting or _____ as a character. How does this author's use of (irony, personification, humor) contribute to _____?
Reading for Information	What does this article tell you about _____? What does the author think about this topic?	What caused this event? In what ways are these ideas important to the topic or theme?	What current event does this remind you of? Does this description fit with what you know about _____?	How useful would this article be for _____? Explain. What could be added to improve the author's argument?
Reading to Perform a Task	What is this supposed to help you do? What time can you get on a non-stop flight to X?	What will be the result of this step in the directions? What must you do before this step?	In order to _____, what information would you need to find that you don't know right now? Describe a situation where you could leave out step X.	Why is this information needed? What would happen if you omitted this?

From: National Center for Education Statistics. (1995). *Reading Assessment Redesigned.* P. 21.

APPENDIX C. Standards for Evaluation of Educational Programs, Projects and Materials

The United States Joint Committee on Standards for Educational Evaluation, which met in 1981 and was composed of representatives from twelve professional education societies, devised standards for evaluating educational programs, projects and materials. These standards can be applied regardless of the size, purpose, or method of the evaluation. In practice, these standards can sometimes conflict, and evaluators usually must balance competing standards. Nonetheless, the existence of evaluation standards and the commitment among evaluators to adhere to them are essential components of the professional field of evaluation.

Utility Standards are intended to help in planning evaluations that are timely, informative and influential and are concerned with whether an evaluation provides practical information to an intended audience.

- ***Audience Identification.*** Audiences involved in or affected by the evaluation should be identified so that their needs can be addressed.
- ***Evaluator Credibility.*** Evaluators should be both trustworthy and competent to perform the evaluation, so their findings achieve maximum credibility and acceptance.
- ***Information Scope and Selection.*** Information collected should address pertinent questions about the object of the evaluation and should be responsive to the needs and interests of designated audiences.
- ***Valuation Interpretation.*** The perspectives, procedures and rationale used to interpret findings should be carefully described so that the bases for value judgements are clear.
- ***Report Clarity.*** The evaluation report should describe the object being evaluated and its context, and the purposes, procedures and findings of the evaluation, so that the audience will readily understand what was done, why it was done, what information was obtained, what conclusions were drawn and what recommendations were made.
- ***Report Dissemination.*** Evaluation findings should be disseminated to clients and other right-to-know audiences so they can assess and use the findings.
- ***Report Timeliness.*** Release of reports should be timely so that audiences can best use reported information.
- ***Evaluation Impact.*** Evaluations should be planned and conducted in ways that encourage follow-through by members of the audiences.

Feasibility Standards deal with issues such as value-for-cost, practical issues such as availability of information and political issues such as impact of the evaluation findings. Feasibility standards require evaluations to be realistic, diplomatic and financially well-managed.

- ***Practical Procedures.*** The evaluation procedures should be practical so that disruption is kept to a minimum and needed information can be obtained.
- ***Political Viability.*** The evaluation should be planned and conducted with anticipation of the different positions of key decision makers.

Propriety Standards are related strongly to ethics and try to ensure that the rights of people participating in and influenced by the evaluation are protected.

- ***Formal Obligation.*** Obligations of the formal parties to an evaluation should be agreed to in writing so that the parties are obliged to adhere to all conditions of the agreement or to formally renegotiate it.
- ***Conflict of Interest.*** Conflict of interest, if unavoidable, should be dealt with openly and honestly so that it does not compromise the evaluation processes and results.
- ***Full and Frank Disclosure.*** Oral and written evaluation reports should be open, direct, and honest in their disclosure of pertinent findings, including evaluation limitations.
- ***Public's Right to Know.*** The formal parties to an evaluation should respect and assure the public's right to know within limits of other related principles and statutes such as those dealing with public safety and the right to privacy.
- ***Rights of Human Subjects.*** Evaluations should be designed and conducted so that the rights and welfare of the human subjects are respected and protected.
- ***Human Interactions.*** Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation.
- ***Balanced Reporting.*** The evaluation should be complete and fair in its presentation of strengths and weaknesses of the object under investigation.
- ***Fiscal Reporting.*** The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and be prudent and ethically responsible.

Accuracy Standards determine whether an evaluation has produced true, valid and reliable knowledge. They require that the data collection and management reflect the key evaluation issues, that the information is technically adequate, and that the conclusions and recommendations that come from the evaluation reflect the analysis of the data.

- ***Object Identification.*** The object of the evaluation should be sufficiently examined so that the form of the object being considered can be clearly identified.
- ***Context Analysis.*** The context in which the program exists should be examined in detail so that its likely influence on the object can be identified.
- ***Described Purposes and Procedures.*** The purposes and procedures of the evaluation should be monitored and described in enough detail to be identified and assessed.
- ***Defensible Information Sources.*** The sources of information should be described in enough detail to allow the adequacy of the information to be assessed.
- ***Valid Measurement.*** The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is valid for the given use.
- ***Reliable Measurement.*** The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is reliable for the given use.
- ***Systematic Data Control.*** The data collected, processed and reported in an evaluation should be reviewed and where necessary corrected, so that the results will not be flawed.
- ***Analysis of Quantitative Information.*** Quantitative information should be appropriately and systematically analyzed to ensure supportable interpretation.
- ***Analysis of Qualitative Information.*** Qualitative information should be appropriately and systematically analyzed to ensure supportable interpretation.
- ***Justified Conclusions.*** The conclusions reached in an evaluation should be explicitly justified so that audiences can assess them.
- ***Objective Reporting.*** The evaluation procedures should provide safeguards to protect the findings against distortion by the personal feelings and biases of any party to the evaluation.

From: The United States Joint Committee on Standards for Educational Evaluation. (1981). *Standards for Evaluation of Educational Programs, Projects and Materials*. Washington, DC: GPO.

Appendix D. Consultants to the Study Group

Dr. Ron Areglado, Associate Executive Director of Programs, National Association of Elementary School Principals

Joan Baratz-Snowden, Deputy Director, Educational Issues, American Federation of Teachers

Dr. David Conley, Associate Professor of Educational Policy, University of Oregon

Dr. Glen Cutlip, Senior Policy Analyst, National Center for Educational Innovation, National Education Association

Dr. Gene Jongsma, Director, Assessment Group, Harcourt Brace

Dr. Stuart Kahl, Vice President, Advanced Systems in Measurement and Evaluation

Dr. Michael Kean, Vice President of Public and Governmental Affairs, CTB/McGraw Hill

Dr. Wayne Martin, Director, State Education Assessment Center, Council of Chief State School Officers

Patty McAllister, Director of the District of Columbia Office, Educational Testing Service

Dr. Linda McGarvey-Levin Director of National Accounts, CTB/McGraw Hill

Monty Neill, Associate Director, FairTest

Dr. Martha Thurlow, Associate Director, National Center on Education Outcomes

Roy Truby, Executive Director, National Assessment Governing Board

Dr. Phil Young, Vice President, Proposals and Contracts Management, Harcourt Brace

Bibliography

August, D., and K. Hakuta. (1997). *Improving Schooling for Language Minority Children: A Research Agenda*. Washington, DC: National Academy Press.

Claycomb, C., and D. Kysilko. (1997). *Issues and Trends in Statewide Assessments*. Alexandria, VA: National Association of State Boards of Education.

CTB/McGraw-Hill. (1996). *Education Assessment: A Primer for School Boards*. Monterey, CA: CTB/McGraw-Hill.

Conley, D. T. (December 1996). "Where's Waldo? The Conspicuous Absence of Higher Education From School Reform and One State's Response." *Phi Delta Kappan*. pp. 309-314.

Conley, D. T. (December 1996). "Daddy, I'm Scared: A Prophetic Parable," *Phi Delta Kappan*. pp. 290-297.

Education Commission of the States. (1997). *So You Have Standards. Now What?* Denver: ECS.

FairTest. (1996). "A New Tidal Wave of Testing, But Still Grounds for Hope." *FairTest Examiner*. v. 10, n. 4, p 16.

Gandal, M. (1997). *Making Standards Matter: An Annual Fifty-State Report on Efforts to Raise Academic Standards*. Washington, DC: AFT.

Harp, L. (July 12,1995). "Kentucky Student Assessments Called 'Seriously Flawed.'" *Education Week*.

Harp, L. (August 14, 1997). "Expand State Tests, End Teacher Bonuses, Panel Urges in Report." *Lexington Herald-Leader*.

Herman, J. L., P. Aschbacher, and L. Winters. (1992). *A Practical Guide to Alternative Assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Jongsma, Gene. (January 24, 1997). "The Current State of Assessments." Presentation to the NASBE Study Group on State Assessment Systems. Alexandria, VA.

Kahl, S. (January 21, 1997). Unpublished comments made at the NASBE Assessment Study Group Advisory Panel Meeting. Alexandria, VA.

Kean, M. (January 24 1997). "Developing a State Assessment System." Presentation to the NASBE Study Group on State Assessment Systems. Alexandria, VA.

Kirst, M., and C. Mazzea. (December 1996). "The Rise, Fall, and Rise of State Assessment in California," *Phi Delta Kappan*. pp. 319-323.

Koretz, D. (Fall 1994). "The Vermont Portfolio Assessment Program: Findings and Implications." *Educational Measurement: Issues and Practice*. v 13, n. 3, Fall. pp. 5-16.

Koretz, D. et al. (1996). *Final Report: Perceived Effects of the Maryland School Performance Assessment Program*. Los Angeles: CRESST/RAND Institute on Education and Training.

Koretz, D. et al. (1996). *Perceived Effects of the Kentucky Instructional Results Information System*. Santa Monica, CA: RAND Institute on Education and Training.

Land, R. (Summer 1997). "Moving Up to Complex Assessment Systems: Proceedings from the 1996 CRESST Conference. *Evaluation Comment*, pp. 1-21.

Linn, R. L., and J. L. Herman. (1997). *A Policymaker's Guide to Standards-Led Assessment*. Denver: Education Commission of the States.

Martin, W. (January 25, 1997). "The Status of New State Assessments: Issues and Trends." Presentation to the NASBE Study Group on State Assessment Systems. Alexandria, VA.

McDonnell, L. M. (1997). *The Politics of State Testing: Implementing New Student Assessments*. CSE Technical Report 424. Los Angeles: National Center for Research on Evaluation, Standards, and Testing.

McDonnell, L. M., M. J. McLaughlin, and P. Morison, eds. (1997). *Educating One & All: Students with Disabilities and Standards-Based Reform*. Washington, DC: National Academy Press.

McGrew, K.S., M. L. Thurlow, J.G. Shriner,, and A.N. Spiegel. (1992). *Inclusion of Students with Disabilities in National and State Data Collection Programs*. Technical Report 2. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

National Center on Educational Outcomes. (1996). *Accommodations for Students with Disabilities: Guidelines from States with Graduation Exams*. State Assessment Series, Minnesota Report #5. Minneapolis: NCEO.

National Center on Educational Outcomes. (1996). *Accommodations for Students with Limited English Proficiency: Analysis of Guidelines from States with Graduation Exams*. Minnesota Report #6. Minneapolis: NCEO.

National Center on Educational Outcomes. (1996). *A Review of the Literature on Testing Accommodations for Students with Disabilities*. State Assessment Series, Minnesota Report #9. Minneapolis: NCEO.

BEST COPY AVAILABLE

National Center on Educational Outcomes. (1997). *Issues and Considerations in Alternate Assessments.* Synthesis Report 27. Minneapolis: NCEO.

National Center for Education Statistics. (1995). *Reading Assessment Redesigned.* Washington, DC: U.S. Department of Education.

National Center for Research on Evaluation, Standards and Student Testing. (1997). "Moving Up To Complex Assessment Systems," *Evaluation Comment.* v. 7, n. 1.

National Education Association. (1995). *The Role of High Stakes Testing in School Reform.* Washington, DC: The National Education Association.

National Education Goals Panel. (1996). *Profile of 1994-95 State Assessment Systems and Reported Results.* Washington, DC: National Education Goals Panel.

National Endowment for the Humanities. (1991). *National Tests: What Other Countries Expect their Students to Know.* Washington, DC: National Endowment for the Humanities.

National Forum on Assessment. (1995). *Principles and Indicators for Student Assessment Systems.* Cambridge, MA: National Center for Fair and Open Testing (FairTest).

Neill, M., et al. (1997). *Testing Our Children: A Report Card on State Assessment Systems.* Cambridge, MA: National Center for Fair and Open Testing (FairTest).

O'Connor, D., and L. Owen. (1996). "Pass or Fail? New Bottom Line for State's Eighth Graders," *St. Paul Pioneer Press.* May 29, p. 1A.

Olson, J. F., and A. A. Goldstein. (1997). *The Inclusion of Students with Disabilities and Limited English Proficient Students in Large-Scale Assessments: A Summary of Recent Progress.* Washington, DC: U.S. Department of Education, National Center for Education Statistics, Office of Educational Research and Improvement.

Olson, L. (March 22,1995). "The New Breed of Assessments Getting Scrutiny," *Education Week.*

Perrone, V., Ed. (1991). *Expanding Student Assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Phillips, S. E. (1995) "Legal Defensibility of Standards: Issues and Policy Perspectives. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES).* Volume II. Washington, DC: NAGB and NCES.

Pitsch, M.. (1996). "State Progress on Title I Rules Called Uneven," *Education Week.* April 10.

Popham, W. J. (1993). "Circumventing the High Costs of Authentic Assessment," *Phi Delta Kappan.*

Resnick, L.B., and D.P Resnick. (1992). "Assessing the Thinking Curriculum: New Tools for Educational Reform." In B.R. Gifford and M.C. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Reville, S.P. (March 27, 1996). "Beyond Public Engagement." *Education Week.*

Rothman, R. (December 1995). "Telling the Story of New Assessments." *The School Administrator.* pp. 6-11.

Smith, M. L. (1996). *The Politics of Assessment: A View from the Political Culture of Arizona.* CSE Technical Report 420. Los Angeles: National Center for Research on Evaluation, Standards, and Testing.

Thurlow, M. (March 15, 1997). Unpublished comments made at the NASBE Assessment Study Group Advisory Panel Meeting. Alexandria, VA.

Thurlow, M., K. Langenfeld, R Nelson, H. Shin, and J. Coleman. (1997). *State Accountability Reports: What Do They Say about Students with Disabilities?* Technical Report 20. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Truby, R. (March 15,1997). Unpublished remarks made at the NASBE Assessment Study Group Meeting. Alexandria, VA.

U.S. Department of Education. (Spring 1996). "What the Research Says about Student Assessment," *Improving America's Schools: A Newsletter on Issues in School Reform.*

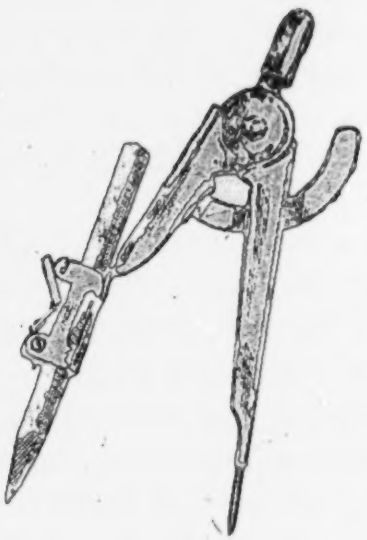
U.S. General Accounting Office. (1994). *Limited English Proficiency: A Growing and Costly Educational Challenge Facing Many School Districts.* GAO/HEHS-94-38.

Ysseldyke, J., *et al.*, (1996). *A Comparison of State Assessment Systems in Kentucky and Maryland.* State Assessment Series Maryland/Kentucky Report 1. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J., M.L. Thurlow, K.S. McGrew, and J.G. Shriner. (1994). *Recommendations for Making Decisions about the Participation of Students with Disabilities in Statewide Assessment Programs.* Synthesis Report 15. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J., K. Olsen, and M. Thurlow. (1997). *Putting Alternate Assessment into Practice: Possible Sources of Data.* Synthesis Report 28. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

The National Association of State Boards of Education is a nonprofit, private association that represents state and territorial boards of education. Our principal objectives are to strengthen state leadership in education policy-making; promote excellence in the education of all students; advocate equality of access to educational opportunity; and assure responsible lay governance of public education.



NASBE
NATIONAL ASSOCIATION OF STATE BOARDS OF EDUCATION

1012 Cameron Street
Alexandria, VA 22314
703-684-4000

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").